YAYI-UIE: A Chat-Enhanced Instruction Tuning Framework for Universal Information Extraction

Anonymous ACL submission

Abstract

The difficulty of the information extraction task lies in dealing with the task-specific label schemas and heterogeneous data structures. Recent work has proposed methods based on large language models to uniformly model different information extraction tasks. However, these existing methods are deficient in their information extraction capabilities for Chinese languages other than English. In this paper, we propose an end-to-end chat-enhanced instruction tuning framework for universal information extraction (YAYI-UIE), which supports both Chinese and English. Specifically, we utilize dialogue data and information extraction data to enhance the information extraction performance jointly. Experimental results show that our proposed framework achieves state-of-the-art performance on Chinese datasets while also achieving comparable performance on English datasets under both supervised settings and zero-shot settings.

1 Introduction

Information extraction (IE) aims to extract structured information from unstructured text automatically (Grishman, 2019). Depending on the extracted objects, the IE tasks can be categorized into multiple sub-tasks, including named entity recognition (NER), relation extraction (RE), event extraction (EE), and so on. The traditional IE methods mostly develop isolated datasets and models for each task, schema and domain, which greatly hinders the practical applications of the IE tasks (Mengge et al., 2020; Wang et al., 2021; Qin et al., 2021; Wang et al., 2022a).

Recently, large language models (LLMs) have demonstrated tremendous capabilities in solving a variety of natural language tasks and are equipped with strong generalization abilities. Therefore, Lu et al. (Lu et al., 2022) first introduced the concept of universal IE to uniformly model various IE tasks. They also proposed a large-scale pre-trained universal IE model called UIE. However, UIE still requires model fine-tuning for different downstream tasks, which leads to its poor performance on unseen data. Lou et al. (Lou et al., 2023) proposed USM and designed three unified token-linking operations to decouple various IE tasks, but its training and inference processes suffer from inefficiency. Wang et al. (Wang et al., 2023) developed an end-to-end unified information extraction framework InstructUIE based on instruction tuning, which utilizes descriptive instructions to enable LLMs to understand different IE tasks. Nevertheless, these existing methods are deficient in their IE capabilities for Chinese languages other than English.

In this paper, we propose YAYI-UIE, an end-to-end chat-enhanced instruction tuning framework for universal information extraction that supports both Chinese and English. Our framework consists of the following two instruction-tuning steps. The first step involves utilizing dialogue data to fine-tune a base LLM for obtaining a chat model with common understanding abilities. In the second step, we focus on enhancing the chat model's performance in IE tasks. To achieve this, we construct the largest and most comprehensive Chinese IE benchmark dataset and combined it with the existing English benchmark. The universal IE model is obtained by instruction-tuning the chat model using this combined dataset.

- We propose an end-to-end instruction tuning framework YAYI-UIE for universal information extraction that supports both Chinese and English, which leverages dialogue data and information extraction data to enhance the information extraction performance jointly.

- We construct the most comprehensive Chinese instruction tuning benchmark for univer-

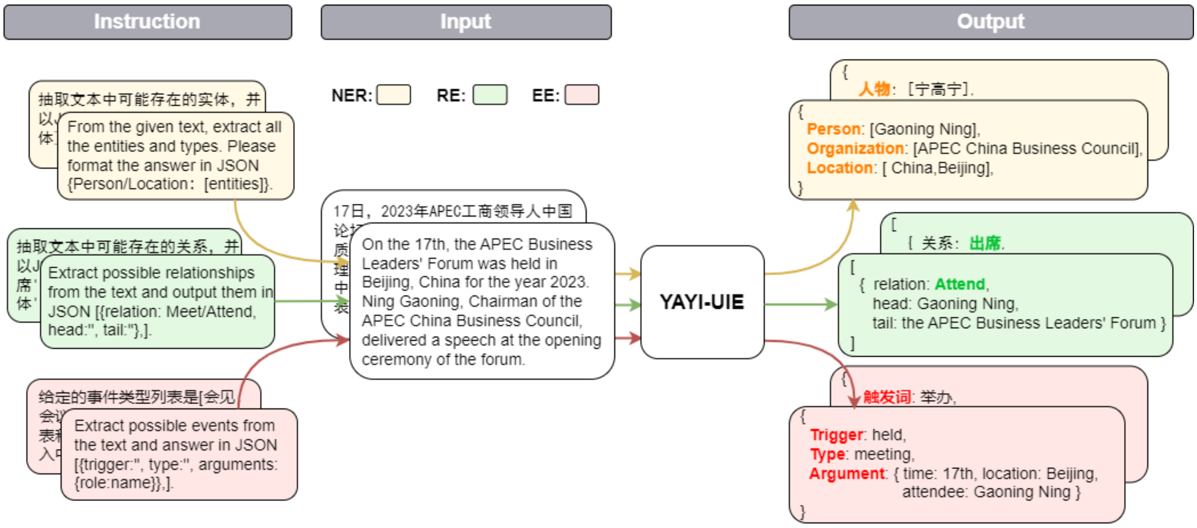

Figure 1: Examples of our chat-enhanced instruction tuning framework for universal information extraction.

sal information extraction, which consists of 16 datasets from various domains.

- The experimental results demonstrate that our YAYI-UIE achieves the SOTA performance in both supervised and zero-shot settings for Chinese, while also displaying remarkable proficiency in English.

2 Methodology

In this section, we describe the proposed chat-enhanced instruction tuning framework for universal information extraction (YAYI-UIE). We start with the task schema and the inference procedure of our universal information extraction framework. Then we introduce the design of the two-step instruction tuning, including instruction tuning for chat and information extraction respectively.

Figure 1 gives examples of our text-to-text generation framework for universal information extraction to illustrate the task schema and the inference procedure. To uniformly model the IE tasks, including NER, RE and EE, we formalize these tasks by the following task schema:

$$\textbf{Ouput} = \textit{YAYI-UIE}\ (\textbf{Instruction}, \textbf{Input}) \quad (1)$$

where the detailed descriptions of the properties in the schema are as follows:

- **Instruction** is a natural language text sequence that includes three elements: task type, task option, and output format. It consists of a description of the task type to specify the task; a description of the task option to restrict the range of the labels in the output; and a description of the desired format of the output.

- **Input** is a textual instance of the IE tasks that is fed to the large language model along with the instruction, and the model generates the output based on the constraints provided by the given instruction.

- **Output** is a sentence that represents the structured information extracted from the input text. Specifically, our YAYI-UIE chooses JSON as the output format for all the IE tasks.

On this basis, we design a two-step instruction tuning for universal information extraction. As shown in Figure 2, we first fine-tune a pre-trained LLM on the dialogue instruction corpus to enhance the instruction-following ability. Following that is the instruction tuning for information extraction, which aims to better constrain the model to generate the desired structured results for the IE tasks.

2.1 Instruction Tuning for Chat

To enhance the model's understanding of open-world languages and improve the performance of instruction fine-tuning in fully supervised and zero-shot settings, intuitively, the dialogue data in real life is a good fit for strengthening the understanding of human language instructions. In the first step of the proposed two-step instruction tuning, we use open-source dialogue data with instructions and a self-constructed corpus to train a chat-enhanced language model to facilitate instruction tuning for multiple information extraction tasks.

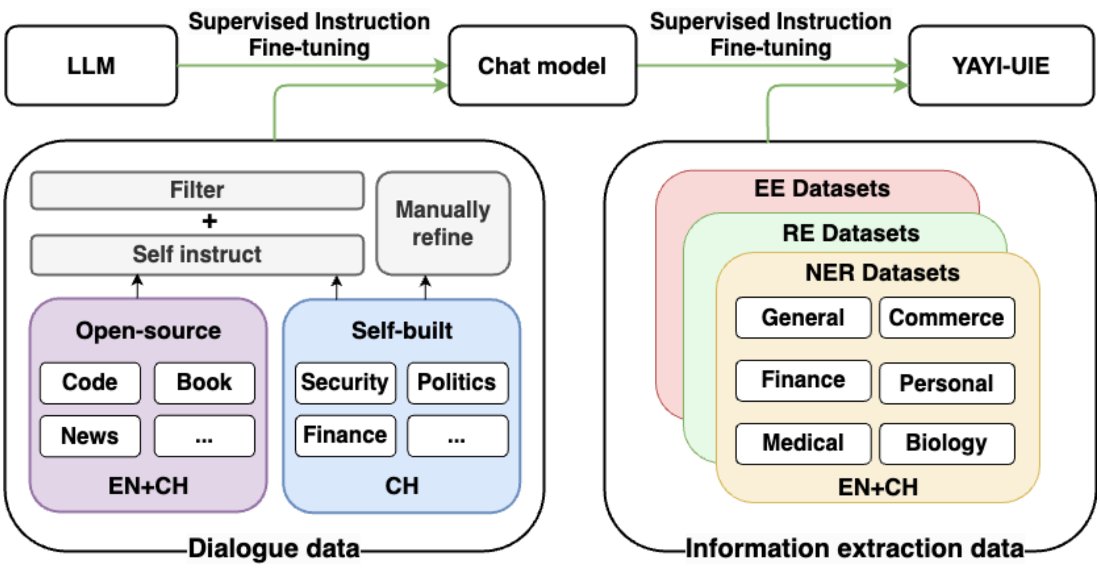

Figure 2: Overview of our chat-enhanced instruction tuning framework for universal information extraction.

Dialogue Data During the data acquisition, to align the model for following human instructions better, we first perform general instruction tuning to train a chat-enhanced model using dialogue corpus in both English and Chinese. The corpus is sourced from the general internet webpages and public datasets, including high-quality data such as news articles, encyclopedic contents, books, codes, etc. In addition to the open-source datasets, we also leverage some field-specified data in the domains of finance, politics, and security. These self-built data, with a large portion in Chinese, include press conference records, company identification, and sensitive boundary recognition, etc.

For data processing, the corpus is constructed based on the self-instruct framework (Wang et al., 2022b), formatted as tuples of instruction, input and output. Specifically, we iteratively perform instruction tuning using the instances generated by our model. At each iteration, the distribution of the generated data is revised using a filtering step, where the meaningless, incomplete, sensitive, or duplicate samples are rejected. In addition, for the field-specified data, we further manually filter the data with regard to format (e.g., line breaks and punctuation errors) and content (e.g., data timeliness and hallucination issues).

Training During the training process, we fine-tuned a base LLM on the constructed dialogue corpus to obtain the chat LLM:

$$\mathbf{LLM}_{chat} = SFT\left(\mathbf{LLM}_{base},\ \mathcal{D}_{dialogue}\right) \quad (2)$$

where $\mathbf{LLM}_{base}$ is a pre-trained LLM, $\mathbf{LLM}_{chat}$ is the fin-tuned chat model, $\mathcal{D}_{dialogue}$ is the constructed dialogue corpus.

2.2 Instruction Tuning for IE

After training on the chat data, the chat model gains a fundamental understanding of open-world language and has been further enhanced in its Chinese language capabilities. In the second step of the proposed two-step instruction tuning, we adapt the model to the IE tasks via the IE instruction datasets and standardize the output format of the model. Moreover, we construct the most comprehensive Chinese IE instruction benchmark dataset to support the supervised fine-tuning for the IE tasks.

Information Extraction Data Due to the lack of Chinese datasets in existing IE benchmarks, we collect 16 Chinese datasets for NER, RE, and EE tasks from diverse domains to build a comprehensive Chinese instruction benchmark, and then combine it with the existing English benchmark IE INSTRUCTIONS (Wang et al., 2023).

Figure 3 gives an overview of the English and Chinese IE data for instruction tuning, which includes the distribution of the data across different tasks, domains, and languages. It covers more than 10 domains, such as general, finance, biology, and healthcare. Specifically, the built IE data covers various label types.

Training To enhance the generalization ability, we perform negative sampling on the labels of each

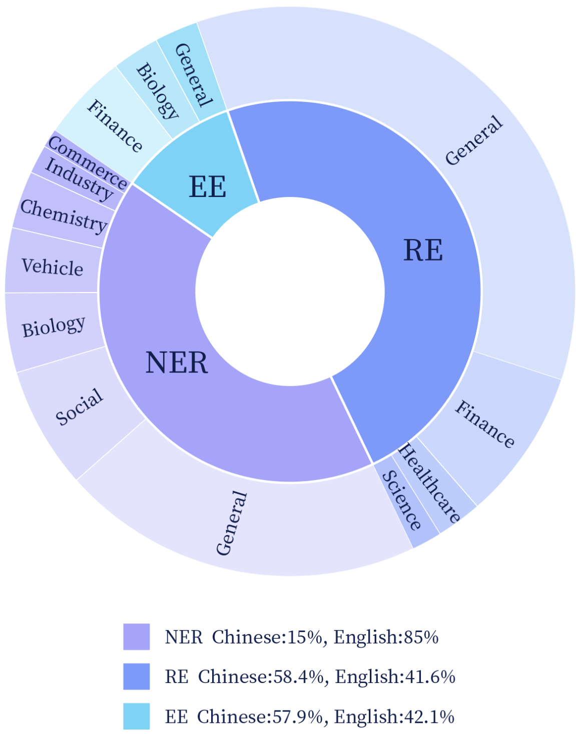

Figure 3: The distribution of information extraction data.

Dataset	BERT-base	UIE	InstructUIE	YAYI-UIE
ACE2005	**87.30**	85.78	86.66	81.78
AnatEM	85.82	77.68	**90.89**	76.54
bc2gm	80.90	74.77	**85.16**	82.05
bc4chemd	86.72	82.79	**90.30**	88.46
bc5cdr	85.28	78.82	**89.59**	83.67
broadtwitter	58.61	67.02	83.14	**83.52**
CoNLL03	92.40	92.99	92.94	**96.77**
FabNER	64.20	73.71	**76.20**	72.63
FindVehicle	87.13	91.56	89.47	**98.47**
GENIA-Ent	73.30	67.46	74.71	**75.21**
HarveyNER	82.26	58.13	**88.79**	69.57
MIT Movie	88.78	79.56	**89.01**	70.14
MIT Rest.	81.02	81.67	**82.55**	79.38
multiNERD	91.25	91.75	**92.32**	88.42
ncbi-disease	80.20	80.13	**90.23**	87.29
Ontonotes	**91.11**	86.25	90.19	87.04
polyglot	**75.65**	68.01	70.15	70.85
tweetNER7	56.49	63.81	64.97	**66.99**
wikiann	70.60	82.11	**85.13**	72.63
wikineural	82.78	**92.14**	91.36	87.63
Avg	80.09	78.81	**85.19**	80.95

Table 1: Overall results of YAYI-UIE on English NER datasets. To provide a comprehensive comparison, we conduct experiments on 18 datasets to obtain the experimental results of UIE, which are marked with underlines.

instance during the training phrase. For input text t containing n types of labels $L = \{l_1, l_2, \cdots, l_n\}$, we randomly add several labels to L that do not belong to L. During the training process, we fine-tuned the chat LLM on the IE corpus to obtain the universal IE model:

$$\mathbf{LLM}_{ie} = \mathit{SFT}\left(\mathbf{LLM}_{chat}, \mathcal{D}_{ie}\right) \tag{3}$$

where $\mathbf{LLM}_{ie}$ is the fine-tuned universal information extraction model, $\mathcal{D}_{ie}$ is the information extraction corpus.

3 Experiments

In this section, we conduct experiments under both supervised settings and zero-shot settings to evaluate the effectiveness of YAYI-UIE. For implementation, we choose Baichuan2-13B (Yang et al., 2023) as the backbone model and perform the proposed chat-enhanced instruction tuning on it with the 10^{-5} learning rate. For the evaluation metrics, we adopt the F1 value to evaluate each dataset in NER, RE and EE tasks in a strict matching manner, and report the respective average F1 of the English datasets and the Chinese datasets on the three tasks.

3.1 Experiments on Supervised Settings

3.1.1 Datasets

We conduct supervised experiments on 32 English datasets and 8 Chinese datasets. The English data provided from the benchmark dataset IE INSTRUCTIONS (Wang et al., 2023). Based on IE INSTRUCTIONS, we further collect 8 Chinese datasets to verify the IE capabilities in Chinese under supervised settings. Specifically, for NER task, we adopt CCKS2017 (Xia and Wang, 2017), CCKS2018 (Luo et al., 2018), MSRA (Levow, 2006), and eCommerce (Liu, 2011) dataset. For RE task, we adopt DuIE (Li et al., 2019) and InstructIE (Gui et al., 2023) dataset. For EE task, we adopt DuEE-Fin (Han et al., 2022) DuEE-1.0 (Li et al., 2020). These datasets cover multiple domains, such as healthcare, finance and biology.

3.1.2 Baselines

We choose the following representative method as the baselines:

- **UIE** (Lu et al., 2022) is a unified text-to-

Dataset	BERT-base	YAYI-UIE
CCKS 2017	**92.68**	90.73
CCKS 2018	**90.82**	90.39
MSRA	**96.72**	95.57
eCommerce	73.70	**88.07**
Avg	88.48	**91.19**

Table 2: Overall results of YAYI-UIE on Chinese NER datasets.

structure generation framework that generates target extraction via schema-based prompts.

- **USM** (Lou et al., 2023) is a unified IE tasks framework, which converts IE tasks to a semantic matching problem.

- **InstructUIE** (Wang et al., 2023) proposes a unified information extraction framework based on multi-task instruction tuning.

- **BERT-base** (Kenton and Toutanova, 2019) refers to task-specific supervised models with state-of-the-art results based on the pretrained language model BERT.

3.1.3 Results

Named Entity Recognition Table 1 gives the experimental results of the comparative methods and YAYI-UIE on 20 English NER datasets. As shown in the table, YAYI-UIE achieves a higher average F1 value than UIE and BERT-base methods. When compared to the strong baseline InstructUIE, YAYI-UIE performs slightly worse. The possible reason is the backbone model and training data of InstructUIE are both limited to English only, while the backbone and fine-tuning data of YAYI-UIE are primarily in Chinese, which may affect its capability on English datasets.

Table 2 gives the experimental results of the comparative methods and YAYI-UIE on 4 Chinese NER datasets. As the existing universal information extraction methods only support the English language, we compare YAYI-UIE to the strong BERT-based methods. In the table, YAYI-UIE achieves the highest average F1 value of 91.19%. For the CCKS 2017, CCKS 2018 and MSRA, YAYI-UIE is only off by less than 2% in F1 values, while for eCommerce, it achieves an improvement of 14.37%. The experimental results show that our model outperforms the baselines on the Chinese NER task.

Dataset	UIE	USM	InstructUIE	YAYI-UIE
ADE corpus	-	-	82.31	**84.14**
CoNLL04	75.00	78.84	78.48	**79.73**
GIDS	-	-	**81.98**	72.36
kbp37	-	-	36.14	**59.35**
NYT	-	-	**90.47**	89.97
NYT11 HRL	-	-	56.06	**57.53**
SciERC	36.53	37.36	**45.15**	40.94
semval RE	-	-	**73.23**	61.02
Avg	-	-	67.98	**68.13**

Table 3: Overall results on English RE datasets.

Dataset	BERT-base	YAYI-UIE
DuIE	74.30	**81.19**
InstructIE	49.21	**59.52**
Avg	61.76	**70.36**

Table 4: Overall results on Chinese RE datasets.

Relation Extraction Table 3 gives the experimental results of the comparative models and YAYI-UIE on 8 English RE datasets. We can see from the table that YAYI-UIE achieves the highest average F1 value, and gains a significant improvement on kbp37 compared to the strong baseline InstructUIE. Compared with UIE and USM, YAYI-UIE performs better on the 2 datasets. It should be noted that the model and code for USM are not available, and we cannot reproduce UIE on these RE datasets due to the lack of position information.

Table 4 gives the experimental results of the comparative models and YAYI-UIE on 2 Chinese RE datasets. From the table, we can see that YAYI-UIE achieves the highest average F1 value, and gains 8.6% F1 improvement compared to the baselines. In general, the experimental results demonstrate the effectiveness of our YAYI-UIE for both English and Chinese RE tasks.

Event Extraction Table 5 gives the Event Trigger and Event Argument F1 value experimental results of the comparative models and YAYI-UIE on 3 English EE datasets. Our YAYI-UIE achieves the highest average F1 score for the event argument extraction task. Compared with UIE, YAYI-UIE performs better on 3 out of 6 datasets, while compared with InstructUIE, YAYI-UIE performs better on 2 out of 6 datasets on EE task.

Table 6 gives the experimental results of the comparative models and YAYI-UIE on 2 Chinese EE

	Dataset	BERT-base	USM	UIE	InstructUIE	YAYI-UIE
Event Trigger	ACE2005	72.5	72.41	73.36	**77.13**	65.00
	CASIE	68.98	**71.73**	69.33	67.80	63.00
	PHEE	-	-	64.77	**70.14**	63.00
	Avg	-	-	69.15	**71.69**	63.67
Event Argument	ACE2005	59.9	55.83	54.79	**72.94**	62.71
	CASIE	60.37	63.26	61.3	63.53	**64.23**
	PHEE	-	-	63.70	62.91	**77.19**
	Avg	-	-	59.93	66.46	**68.04**

Table 5: Overall results on English EE datasets. The results marked with underlines are reproduced in this paper.

	Dataset	UIE	YAYI-UIE
Event Trigger	DuEE-Fin	**84.53**	82.50
	DuEE-1.0	82.18	**85.00**
	Avg	83.36	**83.75**
Event Argument	DuEE-Fin	**75.73**	70.02
	DuEE-1.0	70.68	**78.08**
	Avg	73.21	**74.05**

Table 6: Overall results on Chinese EE datasets.

datasets. In the table, we can see that YAYI-UIE achieves the highest average F1 score for both the event trigger and argument extraction tasks. The experimental results demonstrate that our model outperforms the comparative models on the Chinese EE task.

3.2 Experiments on Zero-shot Settings

3.2.1 Datasets

To validate the zero-shot capability of YAYI-UIE, we collected 16 datasets and tested their performance on three tasks separately, which do not appear in the training set. For NER task, we evaluate English capability on five CrossNER (Liu et al., 2021) subsets (AI, literature, music, politics, science) and Chinese capability on the datasets of boson [1], clue (Xu et al., 2020) and weibo (Peng and Dredze, 2015). For RE task, we evaluate the model's English ability on FewRel (Han et al., 2018) and Wiki-ZSL (Chen and Li, 2021), and Chinese capability on SKE 2020 [2], COAE 2016 [3], IPRE (Wang et al., 2019). For EE task, we test the event argument and event trigger extraction separately, which use Commodity News Corpus (Lee et al., 2022) for the English capability, FewFC (Zhou et al., 2021) and CCF law [4] for the Chinese capability.

3.2.2 Baselines

We choose the following representative models for comparative baselines:

- **ZETT** (Kim et al., 2022) is a framework that extracts relation triplets from unstructured text.
- **ChatGPT** (Ouyang et al., 2022) is a state-of-the-art conversational AI language model that is built upon the GPT-3.5 architecture.
- **ChatGLM** (Du et al., 2022) is an open-source, Chinese-English bilingual conversation language model.
- **KnowLM** (Zhang et al., 2023) is an open-source and extensible knowledge graph extraction tool that can extract entities and relations.

3.2.3 Results

Named Entity Recognition Table 7 gives the zero-shot experimental results of the comparative models and YAYI-UIE on 5 unseen English NER datasets and 3 unseen Chinese NER datasets. For the English NER task, YAYI-UIE outperforms several strong baselines except for ChatGPT. For the Chinese NER task, YAYI-UIE achieves the highest average F1 score.

Relation Extraction Table 8 gives the zero-shot experimental results of the comparative models and YAYI-UIE on 2 unseen English RE datasets, and 3 unseen Chinese RE datasets. We can observe that YAYI-UIE achieves the SOTA on both English and Chinese datasets. For the English RE

[1] https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/boson

[2] https://aistudio.baidu.com/datasetdetail/177191

[3] https://github.com/Sewens/COAE2016

[4] https://aistudio.baidu.com/projectdetail/4201483

Method	EN						CH			
	AI	Literature	Music	Politics	Science	Avg	boson	clue	weibo	Avg
ChatGPT	**54.40**	54.07	**61.24**	**59.12**	**63.00**	**58.37**	38.53	25.44	29.3	31.09
ChatGLM2-6b	0.01	0.03	0.00	0.46	0.68	0.24	1.13	0.07	8.09	3.10
UIE	31.14	38.97	33.91	46.28	41.56	38.37	40.64	34.91	**40.79**	38.78
USM	28.18	**56.00**	44.93	36.10	44.09	41.86	-	-	-	-
InstructUIE	49.00	47.21	53.16	48.15	49.30	49.36	-	-	-	-
KnowLM	13.76	20.18	14.78	33.86	9.19	18.35	25.96	4.44	25.20	18.53
YAYI-UIE	52.40	45.99	51.20	51.82	50.53	50.39	**49.25**	**36.46**	36.78	**40.83**

Table 7: Zero-shot performance on NER task, including 5 English datasets and 3 Chinese datasets.

Method	EN			CH			
	FewRel	Wiki-ZSL	Avg	SKE 2020	COAE2016	IPRE	Avg
gpt-3.5-turbo	9.96	13.14	11.55	24.47	19.31	6.73	16.84
ZETT(T5-small)	30.53	31.74	31.14	-	-	-	-
ZETT(T5-base)	33.71	31.17	32.44	-	-	-	-
InstructUIE	**39.55**	35.20	37.38	-	-	-	-
KnowLM	17.46	15.33	16.40	0.40	6.56	9.75	5.57
YAYI-UIE	36.09	**41.07**	**38.58**	**70.8**	**19.97**	**22.97**	**37.91**

Table 8: Zero-shot performance on RE task, including 2 English datasets and 3 Chinese datasets.

task, YAYI-UIE outperforms the best comparative model InstructUIE in average F1 score by 1.2%. The performance on FewRel is not obviously due to the small size and insufficient learning of the model. For the Chinese RE task, our proposed model performs much better than the baselines.

Event Extraction Table 9 gives the zero-shot experimental results of the comparative models and YAYI-UIE on 1 unseen English EE dataset and 2 unseen Chinese EE datasets. The result shows that YAYI-UIE achieves the SOTA performance for Chinese EE task, and also comparable performance for English EE task. It is worth mentioning that InstructUIE only has English capability, and our model has added a large amount of Chinese data in the training, which has reduced the English capability of the model to some extent.

4 Ablation Study

We conduct the ablation study to further evaluate the effectiveness of the instruction tuning for chat using dialogue data in our framework. We conduct separate experiments with Baichuan2-13B-base and Baichuan2-13B-chat (Yang et al., 2023) as the backbone model for IE instruction fine-tuning. The performances of the two models are measured by calculating the strict F1 score for each dataset. We report the average F1 score for each task. Table 10 shows that Baichuan2-chat's performance for each task significantly outperforms Baichuan2-13B-chat by more than 10 points, which verifies the effectiveness of the chat fine-tuning for the universal information extraction task.

5 Related Work

Large Language Models The advent of large language models (LLMs) has instigated a revolutionary paradigm shift within the field of natural language processing (Guo et al., 2023; Qin et al., 2023; Bubeck et al., 2023). LLMs, such as LLaMA (Touvron et al., 2023a,b), ChatGPT (Ouyang et al., 2022) and GPT4 (OpenAI, 2023), have exhibited remarkable abilities across various applications. These LLMs undergo three primary training stages: pre-training, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF). During the pre-training phase, LLMs gain extensive skills and knowledge. However, they face a challenge in adhering to specific instructions. To mitigate this limitation, SFT is incorporated as a supplementary step. This process entails additional training of the LLM utilizing a dedicated annotated dataset that includes instructions and corresponding responses, augmenting its capabilities in accurately following instructions.

	Method	EN	CH		
		commodity news	FewFC	CCF law	Avg
Event Trigger	ChatGPT	1.41	16.15	0.00	8.08
	UIE	-	50.23	2.16	26.20
	InstructUIE	**23.26**	-	-	-
	YAYI-UIE	12.45	**81.28**	**12.87**	**47.08**
Event Argument	ChatGPT	8.60	44.40	44.57	44.49
	UIE	-	43.02	**60.85**	51.94
	InstructUIE	**21.78**	-	-	-
	YAYI-UIE	19.74	**63.06**	59.42	**61.24**

Table 9: Zero-shot performance on EE task, including 1 English datasets and 2 Chinese datasets.

Task	EN		CH	
	Baichuan-base	Baichuan-chat	Baichuan-base	Baichuan-chat
NER	67.21	**81.21**	66.98	**89.15**
RE	48.99	**65.78**	44.45	**62.67**
Event Argument	44.44	**63.48**	63.03	**68.98**
Event Trigger	45.67	**61.33**	74.31	**84.50**
Avg	51.58	**67.95**	62.19	**76.34**

Table 10: Baichuan-13B-chat and Baichuan-13B-base's performance on each IE task

RLHF, by incorporating human feedback into the training loop, serves as a pivotal mechanism for steering LLMs toward generating high-quality and harmless responses.

Information Extraction Information extraction constitutes a longstanding field devoted to the automated extraction of diverse information structures from unstructured textual sources. Classic IE methods (Mengge et al., 2020; Wang et al., 2021; Qin et al., 2021; Wang et al., 2022a) necessitate the formulation of task-specific architectures and the training of dedicated models, which reveals limitations in the generalization ability of models across diverse IE tasks and imposes stringent demands for annotated data. To fulfill the personalized demands of real-world users. Jiao et al. (Jiao et al., 2023) proposed on-demand IE and developed ODIE to extract the desired content which can be specified by the user. Lu et al. (Lu et al., 2023) also proposed Open-world IE for a more general situation providing broader applicability for information extraction. Lu et al. (Lu et al., 2022) have recently pioneered UIE by uniformly modeling IE tasks with a text-to-structure framework. However, a notable limitation of UIE lies in its deficiency in transferring learning capabilities across diverse tasks or schemas. Lou et al. (Lou et al., 2023) proposed USM by designing three directed token-linking operations to decouple task-specific IE tasks into two extraction abilities, resulting in a notable increase in both training and inference time. Wang et al. (Wang et al., 2023) proposed InstructUIE by utilizing instructive guidance to direct LLMs toward the task, facilitating the generation of target structures. Unfortunately, this method is deficient in IE capabilities for Chinese languages other than English. In this paper, we propose an end-to-end framework YAYI-UIE for universal information extraction that supports both Chinese and English.

6 Conclusion

In this paper, we propose a chat-enhanced instruction tuning framework YAYI-UIE for universal information extraction, and build the most comprehensive Chinese IE instruction benchmark. The proposed framework consists of two instruction-tuning steps. It first utilizes dialogue data to fine-tune a base LLM for obtaining common understanding abilities, and then utilizes the constructed Chinese IE benchmark dataset along with the existing English benchmark for IE instruction fine-tuning. Experimental results show that our proposed framework achieves state-of-the-art performance in Chinese while maintaining English language capabilities.

Limitations

The limitations of our YAYI-UIE are as follows:

- In our experiments, we only choose Baichuan2-13B (Yang et al., 2023) as the backbone model, so the performances of other pre-trained LLMs are not clear.
- In terms of instruction diversity, our training data only includes fewer than 5 types of instruction for each task.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Honghao Gui, Jintian Zhang, Hongbin Ye, and Ningyu Zhang. 2023. Instructie: A chinese instruction-based information extraction dataset. *arXiv preprint arXiv:2305.11527*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 172–183. Springer.

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*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. *arXiv preprint arXiv:2310.16040*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Bosung Kim, Hayate Iso, Nikita Bhutani, Estevam Hruschka, and Ndapa Nakashole. 2022. Zero-shot triplet extraction by template infilling. *arXiv preprint arXiv:2212.10708*.

Meisin Lee, Lay-Ki Soon, Eu Gene Siew, and Ly Fie Sugianto. 2022. CrudeOilNews: An annotated crude oil news corpus for event extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 465–479, Marseille, France. European Language Resources Association.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019. Duie: A large-scale chinese dataset for information extraction. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 791–800. Springer.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 534–545. Springer.

Zhi Liu. 2011. Amazon Commerce reviews set. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55C88.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. *arXiv preprint arXiv:2301.03282*.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. Pivoine: Instruction tuning for open-world information extraction. *arXiv preprint arXiv:2305.14898*.

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)*, pages 5755–5772.

Ling Luo, Nan Li, Shuaichi Li, Zhihao Yang, and Hongfei Lin. 2018. DUTIR at the CCKS-2018 task1: A neural network ensemble approach for chinese clinical named entity recognition. In *CEUR Workshop Proceedings*, volume 2242, pages 7–12.

Xue Mengge, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020. Coarse-to-Fine Pre-training for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6345–6354. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haitao Wang, Zhengqiu He, Jin Ma, Wenliang Chen, and Min Zhang. 2019. Ipre: a dataset for inter-personal relationship extraction. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 103–115. Springer.

Xiao Wang, Shihan Dou, Limao Xiong, Yicheng Zou, Qi Zhang, Tao Gui, Liang Qiao, Zhanzhan Cheng, and Xuanjing Huang. 2022a. MINER: Improving out-of-vocabulary named entity recognition from an information theoretic perspective. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5590–5600. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. InstructUIE: Multi-task instruction tuning for unified information extraction.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-Instruct: Aligning language model with self generated instructions.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* *and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6283–6297. Association for Computational Linguistics.

Yuhang Xia and Qi Wang. 2017. Clinical named entity recognition: Ecust in the ccks-2017 shared task 2. In *CEUR workshop proceedings*, volume 1976, pages 43–48.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

Ningyu Zhang, Jintian Zhang, Xiaohan Wang, Honghao Gui, Kangwei Liu, Yinuo Jiang, Xiang Chen, Shengyu Mao, Shuofei Qiao, Yuqi Zhu, Zhen Bi, Jing Chen, Xiaozhuan Liang, Yixin Ou, Runnan Fang, Zekun Xi, Xin Xu, Lei Li, Peng Wang, Mengru Wang, Yunzhi Yao, Bozhong Tian, Yin Fang, Guozhou Zheng, and Huajun Chen. 2023. Knowlm technical report.

Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li. 2021. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14638–14646.