ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT

Xiang Wei[1], **Xingyu Cui**[1], **Ning Cheng**[1], **Xiaobin Wang**[2], **Xin Zhang, Shen Huang**[2], **Pengjun Xie**[2], **Jinan Xu**[1], **Yufeng Chen**[1], **Meishan Zhang, Yong Jiang**[2], **and Wenjuan Han**[1] ✉
[1] Beijing Jiaotong University, Beijing, China
[2] DAMO Academy, Alibaba Group, China

Abstract

Zero-shot Information Extraction (IE) aims to build IE systems from the unannotated text. This is a challenging task as it involves little human intervention, but it is also worthwhile, as zero-shot IE reduces the time and effort needed for data labeling. Recent research on Large Language Models (LLMs), such as GPT-3 and ChatGPT, has shown promising performance on zero-shot settings. This has inspired us to explore prompt-based methods. In this work, we are the first to quantitatively explore whether strong IE models can be constructed by directly prompting LLMs. Specifically, we transform the zero-shot IE task into a multi-turn question-answering problem with a two-stage framework (namely, ChatIE). With the power of ChatGPT, we extensively evaluate our framework on three IE tasks: entity-relation triple extract, named entity recognition, and event extraction. Empirical results on six datasets across two languages show that ChatIE achieves impressive performance and even surpasses some full-shot models on several datasets (*e.g.*, NYT11-HRL). We believe that our work could shed light on building IE models with limited resources.

1 Introduction

Information extraction aims to extract structured information from unstructured text into structured data formats, including tasks such as entity-relation triple extract (RE), named entity recognition (NER), event extraction (EE) (Ratinov and Roth, 2009; Wei et al., 2020a; Zheng et al., 2021; Li et al., 2020a), *etc*. It is a fundamental and crucial task in natural language processing (Sarawagi et al., 2008). Working with an enormous amount of labeling data is always hectic, labor-intensive, and time-consuming. Hence, many organizations and companies rely on IE techniques to automate manual work with zero/few-shot methods, *e.g.*, clinical IE (Agrawal et al., 2022).

Recent works (Agrawal et al., 2022; Jeblick et al., 2023; Zhang et al., 2022) on large language models (LLMs), such as GPT-3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022) and ChatGPT[1], suggest that LLMs perform well in various downstream tasks even without tuning the parameters but only with a few examples as instructions, but there has been little work investigating their potential for zero-shot IE. Thus, there is a timely question: Is it possible to prompt LLMs to do zero-shot IE tasks under a unified framework? Zero-shot IE tasks are challenging because the structured data containing multiple dependent elements are difficult to extract through one-time prediction, especially for some complex tasks like RE. Previous works decompose these complex tasks into different parts and train several modules to solve each part. For example, in the RE task, the pipeline method PURE (Zhong and Chen, 2021) first identifies two entities and then predicts the relation between them. However, supervision from labeled data is required for this model. Additionally, Li et al. (2019b) regard RE as a question-answering process by first extracting subjects and then objects according to the relation templates.

Based on these clues, in this paper, we turn to ChatGPT and hypothesize that ChatGPT is born with the ability to deposit a unified zero-shot IE model in an interactive mode. More specifically, we propose ChatIE[2] by transforming the zero-shot IE task into a multi-turn question-answering problem with a two-stage framework. In the first stage, we aim to find out the corresponding element types that may exist in a sentence. Then in the second stage, we perform a chain-styled IE to each element type from the first stage. Each stage is implemented with a multi-turn QA process. In each turn, we construct prompts based on designed templates

方法原理

[1] https://openai.com/blog/chatgpt.
[2] Vanilla Prompt vs. ChatIE

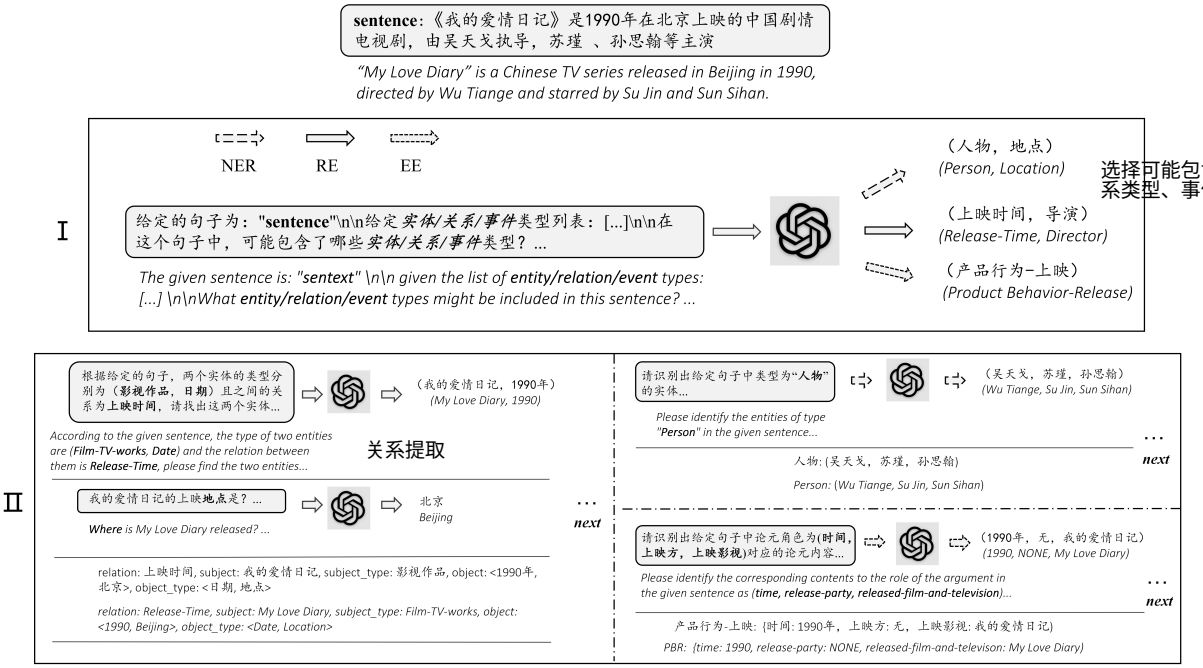

Figure 1: Illustration of the framework. For convenience, we use the samples of DuIE2.0 as examples of three tasks to show.

along with previously extracted information as input to consult ChatGPT. Finally, we compose the information extracted from each turn into the final structured data. We conduct extensive experiments on RE, NER, and EE tasks, including six datasets across two languages: English and Chinese. Empirical results show that while vanilla ChatGPT without using ChatIE fails in solving IE with original task instruction, our proposed two-stage framework instantiated on ChatGPT succeeds when the IE task is decomposed into multiple simpler sub-tasks. Surprisingly, ChatIE achieves impressive performance and even surpasses some full-shot models on several datasets.

2 ChatIE

2.1 Multi-Turn QA framework for zero-shot IE

We introduce the two-stage framework. IE is decomposed into two stages, each containing several turns of QA, which refer to the dialogue with ChatGPT. In the first stage, we aim to find out the existing types of entities, relations, or events in the sentence. In this way, we filter out the element types that do not exist to reduce the search space and computational complexity. Then in the second stage, we further extract relevant information based on the element types extracted in the first stage as well as the corresponding task-specific scheme. The overview of our framework is shown in Fig. 1, which we will describe in detail later.

Stage I: In order to find the element types presented in the sentence, we use one turn of QA with the task-specific template and the list of element types to construct the question. Then we combine the question and sentence as input to ChatGPT. To facilitate answer extraction, we ask the system to reply in the list form. If the sentence does not contain any element types, the system will generate a response of NONE.

Stage II: This stage generally includes multiple QA turns to extract the element for each element type. In advance, we design a series of task-specific question templates for each element type. For complicated schemes such as complex object extraction[3] in entity-relation triple extraction, the length of the chain is greater than one. The extraction of an element may depend on previous elements, so we call it chained templates. We perform multi-turn QA in the order of previously extracted element types as well as the order of ChainExtractionTemplates. To generate a question, we need to retrieve the template according to the element type and fill the corresponding slots if necessary. Then we access ChatGPT and get a response. Finally, we compose structured information based on the elements extracted in each turn. Similarly, for the convenience of answer extraction, we ask the sys-

[3]The complex object refers to an object with multiple attributes.

tem to reply in table form. If nothing is extracted, the system will generate a response with NONE.

2.2 Applying the Framework to IE tasks

After curating the unified framework, we'll then apply the framework to IE tasks, to process and build models for each task.

2.2.1 Entity-Relation Triple Extraction

Given a sentence x and question prompt $q = \{q_1, q_2, ...\}$, the model is desired to predict triples $T(x) = \{(s_1, r_1, o_1), \cdots, (s_n, r_n, o_n)\}$, where $type((s_i, r_i, o_i)) \in \mathcal{T}$. $\mathcal{T}$ denotes the list of potential triple types. Formally for an output triple (s, r, o), we can express the process as:

$$p((s,r,o)|x,q) = \underbrace{p(r|x,q_1)}_{\text{Stage I}} \underbrace{p((s,o)|q_2) \overbrace{\cdots\cdots}^{\substack{complex \quad object}}}_{\text{Stage II}} \quad (1)$$

where q_1 is the question generated using relation types list R and the corresponding template in Stage I. And q_2 in Stage II is the question generated using the template related to the previously extracted relation type. It is worth noting that we have not explicitly shown x in Stage II terms, but ChatGPT can record the relevant information of each turn QA. In addition, we need several further turns QA for samples with complex objects.

2.2.2 Named Entity Recognition

For the NER task, Stage I is to filter out the existing entity types in the sentence given the desired type list. Once we get the entity types, we can construct the input for the second stage accordingly. In Stage II, each turn aims to extract the entities of one type. So the number of turns in Stage II is up to the number of entities obtained in Stage I, and Stage II is omitted if the first stage gets no types at all.

实体标注两阶段

2.2.3 Event Extraction

ChatIE divides the zero-shot EE task into two subtasks: event classification and argument extraction. Stage I is designed for event classification. We formalize it as a classification problem to obtain event types for a given text. Stage II is then devoted to argument extraction. We formalize it as an extractive machine read comprehension problem that identifies arguments of specific roles associated with predicted event types from Stage I.

3 Experiment

3.1 Datasets and Baselines

We experiment on six datasets (Appendix A) in Chinese and English (Tab.1). For each dataset, we provide few-shot baseline models (*i.e.*, **Row fs-1/5/20/100**) as well as full-shot baseline models (*i.e.*, **Row full-shot**) with the same model architecture: PaddleNLP LIC2021 IE[4], CasRel (Wei et al., 2020a), AdaSeq Bert-CRF[5], AdaSeq Bert-CRF, PaddleNLP LIC2021 EE[6], Text2Event-T5-base (Lu et al., 2021) for DuIE2.0 (Li et al., 2019a), NYT11-HRL (Takanobu et al., 2019), MSRA (Levow, 2006), conllpp (Wang et al., 2019), DuEE1.0 (Li et al., 2020c), and ACE05[7], respectively. We also provide a zero-shot baseline (*i.e.*, **Row zs-uie**) UIE (Lu et al., 2022), a universal SOTA IE model. Although supervised approaches and zero-shot approaches are incomparable, we provide the results of SOTA supervised approaches (*i.e.*, **Row Sup-SOTA**) for reference only: HIKNLU[8], RERE (Xie et al., 2021), BERT-MRC+DSC (Li et al., 2020b), Noise-robust Co-regularization + LUKE(Zhou and Chen, 2021), EEQA (Du and Cardie, 2020), HIKNLU for DuIE2.0, NYT11-HRL, MSRA, conllpp, DuEE1.0, and ACE05, respectively. We provided the reported scores. For those unreported results, we re-implement the model and train it three times to obtain an average result. We randomly select exemplars for few-shot settings.

3.2 Evaluation Metrics

RE. We report the standard micro F1 measure and adopt two evaluate metrics (following Zhong and Chen (2021)): *border* evaluation (Rel) and *strict* evaluation (Rel+, appendix B). We use Rel on NYT11-HRL because there is no annotation of entity types and use Rel+ on DuIE2.0.

NER. We consider the complete matching and use the micro F1. Only when both the boundary and the type of the predicted entity are correct, will we regard it as correct.

EE. We adopt different evaluation metrics on the DuEE1.0 and ACE05 dataset. For the DuEE1.0 dataset, F-measure (F1[6]) is scored according to the word-level matching. For the ACE05 dataset, the predicted argument results are matched with

[4] github.com/PaddlePaddle/PaddleNLP/tree/develop/examples/information_extraction/DuIE. The default model is ernie-3.0-medium-zh

[5] github.com/modelscope/AdaSeq/tree/master/examples/bert_crf

[6] github.com/PaddlePaddle/PaddleNLP/tree/develop/examples/information_extraction/DuEE default model is ernie-3.0-medium-zh

[7] https://catalog.ldc.upenn.edu/LDC2006T06

[8] https://aistudio.baidu.com/aistudio/competition/detail/46/0/leaderboard

	RE						NER						EE					
	DuIE2.0#			NYT11-HRL			MSRA#			conllpp			DuEE1.0#			ACE05		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
zs-uie	-	-	0.0	-	-	0.3	-	-	35.21	-	-	13.73	-	-	0.0	-	-	0.25
fs-1	0.0	0.0	0.0	0.0	0.0	0.0	14.7	7.9	9.7	2.71	17.2	4.66	0.4	0.2	0.3	0.0	0.0	0.0
fs-5	0.0	0.0	0.0	0.0	0.0	0.0	34.5	10.3	15.5	2.53	16.65	4.38	0.2	0.6	0.3	0.0	0.0	0.0
fs-20	41.4	0.4	0.8	3.4	2.7	0.5	63.4	44.8	52.5	2.48	19.36	4.41	1.7	0.8	1.1	4.6	0.1	0.2
fs-100	50.8	7.2	12.0	34.8	6.2	10.6	81.3	76.1	78.6	50.26	24.97	32.89	8.7	12.0	10.1	8.0	4.9	6.0
full-shot	68.9	72.2	70.5	47.88*	55.13*	51.25*	96.33	95.63	95.98	94.18	94.61	94.39	50.9	42.8	46.5	45.3	54.3	49.4
Single	17.8	7.7	10.7	10.8	5.7	7.4	55.4	52.2	53.7	**61.2**	42.0	49.8	61.7	77.5	68.7	10.8	16.9	13.2
ChatIE	**74.6**	**67.5**	**70.9**	**30.6**	**48.4**	**37.5**	**58.7**	**53.2**	**55.8**	59.7	**47.5**	**52.9**	**66.5**	**78.5**	**72.0**	**11.6**	**18.5**	**14.3**
Single-api	7.41	14.09	9.71	4.67	11.62	6.61	38.74	55.06	45.48	56.78	69.65	62.56	49.06	63.45	55.33	9.35	16.39	11.91
ChatIE-api	49.94	50.67	50.31	16.58	26.76	20.48	50.22	64.06	56.30	72.62	58.86	65.02	48.49	69.63	57.17	12.24	19.89	15.15
Sup-SOTA	82.44*	80.68*	81.55*	52.40*	58.91*	55.47*	-	-	96.7*	-	-	95.88*	86.02*	84.41*	85.21*	-	-	63.9*

Table 1: F1 score on six datasets over two languages, # denote Chinese. * denote reported scores. Sup: SOTA supervised approaches. Details about the results refer to Appendix C.

the manually marked argument results at the entity level and evaluated by the micro F1.

4 Results

We summarize the main results in Tab. 1[9]. We observe that while the baseline model (**Row Single**; ChatGPT using a single-turn QA instead of ChatIE) performs poorly in solving IE, our proposed two-stage framework based on ChatGPT (**Row ChatIE**) succeeds. ChatIE generally improves performance over six widely used IE datasets by 16.65% points significantly on average. In addition, we have surpassed zero-shot UIE (**Row zs-uie**) in every way.

Notably, the gains become more significant compared with few-shot approaches (**Row fs-**·) even though ChatIE is zero-shot setting. ChatIE is comparable to fs-20 on MSRA, and outperforms fs-100 on NYT11-HRL, conllpp, and ACE05.

More surprisingly, ChatIE even surpasses the full-shot models (Row **full-shot**) on DuIE2.0 and DuEE1.0 even though they are independently trained from scratch using high-quality labeled data. Moreover, compared the supervised model MultiR (Hoffmann et al., 2011) with F1 score 31.7% on NYT11-HRL, ChatIE surpassed it by 5.8%.

Recent work Chen et al. (2023) showed that ChatGPT has worsened over the months. To verify whether this has an impact on our approach, we experimented with `gpt-3.5-turbo-0301` (**Row *-api**) following Kojima et al. (2022). The results show that our method is still highly superior.

In addition, to showcase ChatIE's applicability to a wide range of LLMs, we have tried to apply ChatIE to other different LLM backbones, including ChatGLM2[10], InstructGPT (Ouyang et al., 2022) and LL2ma2-7b-chat[11]. The results are shown in Tab. 2. We can observe that across the different LLM backbones, our framework is still valid. `Multi` refers to applying our multi-round IE framework. `Single` is the baseline approach with only one round of QA.

	ChatGLM2	InstructGPT	LLama2-7b-chat
Single	19.60	9.75	6.65
Multi	22.01	29.31	10.15

Table 2: F1 results on other LLMs with different backbones.

5 Analysis

Robustness. We conduct experiments[12] to analyze the impact of different prompts. The experimental data consisted of 100 randomly sampled samples and the results are shown in Tab.4. We can find that the variance of F1 is very small, indicating that changes due to different wording and phrasing in the textual prompts do not have a huge impact on performance. Thus it shows the robustness of our method.

Data Leakage. Data leakage during model evaluation occurs when data from the training set passes into the test set. This data leakage causes the model's performance estimate on the test set to be biased. LLMs are trained using extremely large data from websites, *etc*. This results in a huge problem, where samples from the test set may leak into the dataset used to train the model.

To address this concern, we prepared three new test datasets that have never been released before. Specifically, we randomly sampled 100 samples from the existing conllpp data and modified it using entity replacement to make sure the samples do

[9]The experiments of Single/ChatIE are conducted using the version of ChatGPT prior to February 9, 2023.

[10]https://github.com/THUDM/ChatGLM2-6B

[11]https://github.com/facebookresearch/llama

[12]using `gpt-3.5-turbo` following Kojima et al. (2022).

Error Type	Percentage(MSRA/conllpp)	Example
I. Correct Boundary but False Type	9.52% / 17.79%	**Sentence**: But China saw their luck desert them in the second match of the group, crashing to a surprise 2-0 defeat to newcomers Uzbekistan. **Expected Output:**["China", "LOC"] **Output:**["China", "GPE"]
II. Correct Type but False Boundary	9.38% / 2.41%	**Sentence**: Physical prices for the weekend at the AECO storage hub were also down about 10 cents in the C$1.92-1.97 per gigajoule, or $1.52-1.56 per mmBtu range, pressured by unseasonably mild weather in western Canada. **Expected Output:**["Canada", "LOC"] **Output:**["western Canada", "LOC"]
III. Unrecognized	54.78% / 56.18%	**Sentence**: The Syrians scored early and then played defensively and adopted long balls which made it hard for us. **Expected Output:**["Syrians", "MISC"] **Output:**[]
IV. Over-recognized	26.34% / 23.63%	**Sentence**: 361 Group A **Expected Output:**[] **Output:**["361 Group A", "MISC"]

Table 3: Error analysis for NER.

No.	Template	F1(%)
1	Please recognize the entities of "" type in the given sentence: ""	45.08
2	Which entities of type "" are contained in the given sentence ""?	44.48
3	In the following sentence "", find the entities with type "".	45.12
4	Knowing the sentence "", identify the entities of type "" in it.	44.78
5	In the given sentence "", the entities of type "" are:	43.77
Average		44.65
Variance		**0.003**(%)

Table 4: Results on NER for different prompts.

	P	R	F1
Original	32.07	82.16	**46.13**
Test I	31.62	80.43	45.40
Test II	31.80	83.06	45.99
Test III	30.62	75.96	43.64
Average	31.34	79.82	**45.01**

Table 5: Analysis of data leakage.

not exist in the original dataset. In terms of entity replacement, we use conllpp test data to collect entities belonging to the same entity type. Then, for each sentence, all the entities are replaced with entities with the same entity type. We manually check the modified sentences to ensure their quality. We build three datasets (*i.e.*, Test I/II/III). We experiment[12] on the three new datasets and find that although a slight decrease is observed compared with the original dataset (from 46.13 to 45.01), ChatIE still achieves an improvement compared with the baseline model. The detailed results are shown in Tab. 5.

6 Case Study

Tab. 6 demonstrates some cases from NYT11-HRL predicted by ChatIE for the IE task. The first sample is an RE case where the same pair of entities belong to two different types of relations. The triples are *(India, location-contains, Delhi)* and *(Delhi, administration_division-country, India)*. In the first stage, ChatIE detects the two relation types. Then in the second stage, ChatIE further extracts *Delhi* and *India*. This shows ChatIE's ability to give different labels to the same entity in different relations. It is worth noting that we convert *location-contains* to *location-located_in* in the experiment and this conversion has not changed the results. It implies that ChatGPT is able to recognize the equivalence of *(Delhi, location-located_in, India)* and *(India, location-contains, Delhi)*.

The second sentence "Four other ***Google*** executives the chief financial officer, ***George Reyes***; the senior vice president for business operations, ***Shona Brown***; the chief legal officer, ***David Drummond***; and the senior vice president for product management, ***Jonathan Rosenberg*** earned salaries of $ 250,000 each." is an RE example where one relation involves multiple triples. It's hard for

many methods to extract all triples but it is accomplished by ChatIE. The extracted triples are *(George Reyes, person-company, Google)*, *(Shona Brown, person-company, Google)*, *(David Drummond, person-company, Google)* and *(Jonathan Rosenberg, person-company, Google)*. ChatIE first filters out the *person-company* type and outputs the 4 triples related to the relation at the same time in the second stage.

The third sentence "Score on the first day of the four-day Sheffield Shield match between ***Tasmania*** and ***Victoria*** at Bellerive Oval on Friday." is a NER example with confusing entities. Both the word *Tasmania* and *Victoria* can be categorized as "LOCATION" types, but they are actually team names in this sentence, which are "ORGANIZATION" types. ChatIE can recognize confusing entities, showing its advantage in understanding ambiguous word senses and choosing the right word sense.

The last sentence "***Clinton*** suffered greatly over the ***19 Rangers*** that ***died***, 18 on the ***3rd of October*** and MattReersen (ph) ***three days later***." is an EE example. In the first stage, ChatIE gets the event type when scanning the word "died". Then it goes from this word to catch the victim "19 rangers", further detects the agent "Clinton" before the predicate, and targets on "3rd of October" and "three days later".

7 Error Analysis

We conduct experiments of the error analysis w.r.t. MSRA and conllpp. We observe that there are mainly four error types as shown in Tab. 3.

- I *Correct Boundary but False Type.* Sometimes, this error type can't be attributed to the capability of the LLM, since the "incorrect" types are reasonable for humans. Take the first column in Tab. 3 as an example, "China" is classified as a "GPE" entity (*i.e.*, geo-political entity), but appeared to be the "LOC" entity as the ground-truth label. "GPE" type is actually reasonable for humans.

- II *Correct Type but False Boundary.* The reason for entity boundary error is kind of complicated. Often, the predicted false boundaries are acceptable and can be explained as different granularity. The percentage of this error type is higher on MSRA, showing the difficulty in word segmentation in Chinese compared with English.

RE: entities belonging to two relations
Just as the JAMA article was being published, three dozen children began dying of acute renal failure at two hospitals in ***Delhi***, ***India***.
RE: one relation involving multiple triples
Four other ***Google*** executives the chief financial officer, ***George Reyes***; the senior vice president for business operations, ***Shona Brown***; the chief legal officer, ***David Drummond***; and the senior vice president for product management, ***Jonathan Rosenberg*** earned salaries of $ 250,000 each.
NER: confusing entities
Score on the first day of the four-day Sheffield Shield match between ***Tasmania*** and ***Victoria*** at Bellerive Oval on Friday.
EE: predicate in a clause
Clinton suffered greatly over the ***19 Rangers*** that ***died***, 18 on the ***3rd of October*** and MattReersen (ph) ***three days later***.

Table 6: Illustration of the case study.

- III *Unrecognized.* The unrecognized errors are mainly due to incomprehension of the sentence, and it is not ruled out that the context of the given sentence is not enough.

- IV *Over-recognized.* This error type is a common error for both datasets, which could be attributed to the ambiguity of the entity type. "361 Group A" is indeed an organization belonging to the "MISC" type. But the "MISC" type is not predefined for MSRA and conllpp. We speculate that this is due to the presence of such a type in the training dataset for LLMs.

8 Prompt of Vanilla Prompt *vs*. ChatIE

Tab. 7, 9 and 8 demonstrate the comparison of vanilla prompts (Row **Single**) and our Chat-based prompts (Row **ChatIE**).[13]

9 Related Work

Working with an enormous amount of labeling data is always hectic, labor-intensive, and time-consuming. Hence, researchers focus on zero/few-shot technologies even though IE is challenging

[13]The experiments are conducted using the version of ChatGPT prior to January 30, 2023.

1	Vanilla Prompt	Chat-based Prompt
STAGE I	***Question:*** Suppose you are an entity-relationship triple extraction model. I'll give you list of head entity types: subject_types, list of tail entity types: object_types, list of relations: relations. Give you a sentence, please extract the subject and object in the sentence based on these three lists, and form a triplet in the form of (subject, relation, object). The given sentence is "Bono said that President Jacques Chirac of France had spoken eloquently of the need to support Africa , though he added that France had not yet come through with the resources ." relations:['location-located_in','administrative_division-country', 'person-place_lived', 'person-company', 'person-nationality', 'company-founders', 'country-administrative_divisions', 'person-children', 'country-capital','deceased_person-place_of_death','neighborhood-neighborhood_of', 'person-place_of_birth'] subject_types: ['organization', 'person', 'location', 'country'] object_types: ['person', 'location', 'country', 'organization', 'city'] In the given sentence, what triples might be contained? Please answer in the form (subject, relation, object): - ***Expected Output:*** [(Jacques Chirac, person-nationality, France)] ***Output:*** []	***Question:*** The given sentence is " Bono said that President Jacques Chirac of France had spoken eloquently of the need to support Africa , though he added that France had not yet come through with the resources ." List of given relations: ['location-located_in','administrative_division-country', 'person-place_lived', 'person-company', 'person-nationality', 'company-founders', 'country-administrative_divisions', 'person-children', 'country-capital','deceased_person-place_of_death','neighborhood-neighborhood_of', 'person-place_of_birth'] What relations in the given list might be included in this given sentence? If not present, answer: none. Respond as a tuple, e.g. (relation 1, relation 2, ...): - ***Expected Output:*** (person-nationality) ***Output:*** (person-nationality)
STAGE II	None	***Question:*** According to the given sentence, the two entities are of type ('person', 'country') and the relation between them is 'person-nationality', find the two entities and list them all by group if there are multiple groups. If not present, answer: none. Respond in the form of a table with two columns and a header of ('person', 'country'): - ***Expected Output:*** (Jacques Chirac, France) ***Output:*** (Jacques Chirac, France)

Table 7: Illustration of vanilla prompts vs our Chat-based prompts in terms of RE. The text highlighted with red represents the prompt template. The text following ***Question:*** represents the prompt that is used in ChatIE.

in low-resource scenarios, such as few-shot relation classification or extraction (Sainz et al., 2021; Han et al., 2018), few-shot event argument extraction (Sainz et al., 2022a) and few-shot information extraction(Sainz et al., 2022b).

ChatGPT has gained widespread attention recently. There are a great many studies w.r.t. downstream NLP tasks. For example, Zhang et al. (2022) leveraged ChatGPT and achieved state-of-the-art performance on Stance Detection. Guo et al. (2023) evaluated its helpfulness in question answering. Jiao et al. (2023) stated that it is a good translator for spoken language. Many other fields also had received its impacts and evolved fast, such as

1	Vanilla Prompt	Chat-based Prompt
STAGE I	***Question:*** The list of argument roles corresponding to the event type 'Contact:Phone-Write' is ['Entity', 'Time'], The list of argument roles corresponding to the event type 'Business:Declare-Bankruptcy' is ['Org', 'Time', 'Place'], The list of argument roles corresponding to the event type 'Justice:Arrest-Jail' is ['Person', 'Agent', 'Crime', 'Time', 'Place'], The list of argument roles corresponding to the event type 'Life:Die' is ['Agent', 'Victim', 'Instrument', 'Time', 'Place'], The list of argument roles corresponding to the event type 'Personnel:Nominate' is ['Person', 'Agent', 'Position', 'Time', 'Place'], The list of argument roles corresponding to the event type 'Conflict:Attack' is ['Attacker', 'Target', 'Instrument', 'Time', 'Place'], The list of argument roles corresponding to the event type 'Justice:Sue' is ['Plaintiff', 'Defendant', 'Adjudicator', 'Crime', 'Time', 'Place'], The list of argument roles corresponding to the event type 'Life:Marry' is ['Person', 'Time', 'Place']. Give a sentence:"What I do know is Saddam Hussein has butchered over a million of his own citizens.", please extract the event arguments according to the argument roles, and return them in the form of a table.The header of the table is 'event type', 'argument role', 'argument content'. If no argument role has a corresponding argument content, the argument content returns "None". ***Expected Output:*** "event_type": "Life:Die", "arguments": ["role": "Victim", "argument": "over a million of his own citizens" , { "role": "Agent", "argument": "Saddam Hussein" } ***Output:*** None	***Question:*** The list of event types: ['Life:Die', 'Justice:Arrest-Jail', 'Contact:Phone-Write', 'Life:Marry', 'Conflict:Attack', 'Personnel:Nominate', 'Business:Declare-Bankruptcy', 'Justice:Sue'] Give a sentence: "What I do know is Saddam Hussein has butchered over a million of his own citizens." What types of events are included in this sentence? Please return the most likely answer according to the list of event types above. Require the answer in the form: Event type ***Expected Output:*** Life:Die ***Output:*** Life:Die
STAGE II	None	***Question:*** The list of argument roles corresponding to the event type 'Life: Die' is ['Agent', 'Victim', 'Instrument', 'Time', 'Place']. please extract the event arguments in the given sentence according to the argument roles, and return them in the form of a table. The header of the table is 'event type', 'argument role', 'argument content'. If no argument role has a corresponding argument content, the argument content returns "None". ***Expected Output:*** "arguments": ["role": "Victim", "argument": "over a million of his own citizens" , { "role": "Agent", "argument": "Saddam Hussein" } ***Output:*** "arguments": ["role": "Victim", "argument": "over a million of his own citizens" , { "role": "Agent", "argument": "Saddam Hussein" }

Table 8: Illustration of vanilla prompts vs our Chat-based prompts in terms of EE. The text highlighted with red represents the prompt template. The text following ***Question:*** represents the prompt that is used in ChatIE.

1	Vanilla Prompt	Chat-based Prompt
STAGE I	***Question:*** I'm going to give you a sentence and ask you to identify the entities and label the entity category. There will only be 4 types of entities: ['LOC', 'MISC', 'ORG', 'PER']. Please present your results in list form. "Japan then laid siege to the Syrian penalty area and had a goal disallowed for offside in the 16th minute." Make the list like: ['entity name1', 'entity type1'],['entity name2', 'entity type2'].....	***Question:*** Given sentence: "Japan then laid siege to the Syrian penalty area and had a goal disallowed for offside in the 16th minute." The known entity types are: ['LOC', 'MISC', 'ORG', 'PER']. Please answer: What types of entities are included in this sentence?
	Expected Output: ["Japan", "LOC"], ["Syrian", "MISC"] ***Output:*** []	***Expected Output:*** LOC, MISC ***Output:*** LOC, MISC
STAGE II	None	***Question:*** According to the sentence above, please output the entities of 'LOC' in the form of list like: ['entity name1', 'entity type1'], ['entity name2', 'entity type2'].....
		According to the sentence above, please output the entities of 'MISC' in the form of list like: ['entity name1', 'entity type1'], ['entity name2', 'entity type2'].....
		Expected Output: ["Japan", "LOC"], ["Syrian", "MISC"]***Output:*** ["Japan", "LOC"], ["Syrian", "LOC"]

Table 9: Illustration of vanilla prompts vs our Chat-based prompts in terms of NER. The text highlighted with red represents the prompt template. The text following ***Question:*** represents the prompt that is used in ChatIE.

Medicine (Jeblick et al., 2023; King, 2022) and Online Exam (Susnjak, 2022). We try to explore its information extraction capabilities and propose a simple but effective zero-shot IE framework.

10 Conclusion

To the best of our knowledge, we quantitatively investigate for the first time whether strong IE models can be constructed by directly prompting LLMs. We presented ChatIE, a multi-turn QA framework for zero-shot information extraction based on ChatGPT. Through this interactive mode, ChatIE can decompose complex IE tasks into several parts and compose the results of each turn into a final structured result. We apply this framework to RE, NER, and EE tasks and conduct extensive experiments on six datasets across two languages to validate its effectiveness. Surprisingly, ChatIE achieves impressive performance and even surpasses some full-shot models on several datasets. This work paves the way for a new paradigm for zero-shot IE, where the experts decompose IE task into multiple simpler and easier sub-tasks, define chat-like prompts, and directly runs those specifications without training and finetuning.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683.

Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the*

2015 Conference on Empirical Methods in Natural Language Processing, pages 1774–1784.

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*.

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.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2023. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *European Radiology*, pages 1–9.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Michael R King. 2022. The future of ai in medicine: a perspective from a chatbot. *Annals of Biomedical Engineering*, pages 1–5.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2021. Unified named entity recognition as word-word relation classification. *ArXiv*, abs/2112.10070.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019a. Duie: A large-scale chinese dataset for information extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 791–800. Springer.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019b. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020c. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 534–545. Springer.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end 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 2795–2806.

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, Dublin, Ireland. Association for Computational Linguistics.

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.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022a. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. 2022b. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 27–38, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Sunita Sarawagi et al. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7072–7079.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. Deepstruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. Crossweigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020a. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020b. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online. Association for Computational Linguistics.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. Revisiting the negative data of distantly supervised relation 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 3572–3581, Online. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. Prgc: Potential relation and global correspondence based joint relational triple 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 6225–6235.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683.

Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784.

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*.

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.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2023. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *European Radiology*, pages 1–9.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Michael R King. 2022. The future of ai in medicine: a perspective from a chatbot. *Annals of Biomedical Engineering*, pages 1–5.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2021. Unified named entity recognition as word-word relation classification. *ArXiv*, abs/2112.10070.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019a. Duie: A large-scale chinese dataset for information extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 791–800. Springer.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019b. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020c. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 534–545. Springer.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end 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 2795–2806.

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, Dublin, Ireland. Association for Computational Linguistics.

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.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022a. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. 2022b. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 27–38, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Sunita Sarawagi et al. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7072–7079.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. Deepstruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. Crossweigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020a. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020b. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online. Association for Computational Linguistics.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. Revisiting the negative data of distantly supervised relation 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 3572–3581, Online. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. Prgc: Potential relation and global correspondence based joint relational triple 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 6225–6235.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

A Details of Data

We experiment on six datasets. For each dataset, we provide few-shot baseline models (*i.e.*, Row fs-1/5/10/20/50/100) as well as full-shot baseline models (*i.e.*, Row full-shot). Although supervised approaches and few-shot approaches are incomparable, we provide the results of SOTA supervised approaches (*i.e.*, Row Sup-SOTA) for reference only.

RE. NYT11-HRL (Takanobu et al., 2019) is a preprocessed version of NYT11 (Riedel et al., 2010; Hoffmann et al., 2011) and contains 12 predefined relation types. DuIE2.0 (Li et al., 2019a) is the industry's largest schema-based Chinese RE dataset and contains 48 predefined relation types[14].

For each few-shot experiment, we train three times on randomly selected sets from the training data to get an average result.

NER. The conllpp (Wang et al., 2019) dataset is a modified version of the conll2003 (Tjong Kim Sang and De Meulder, 2003) and contains 4 entity types. MSRA (Levow, 2006) is a Chinese named entity recognition dataset for the news field and contains 3 entity types.

EE. DuEE1.0 (Li et al., 2020c) is a Chinese event extraction dataset released by Baidu, which contains 65 event types. The ACE05[15] corpus provides event annotations in document and sentence levels from a variety of domains such as newswires and online forums.

[14]The dataset not specifically specified is an English dataset.

[15]https://catalog.ldc.upenn.edu/LDC2006T06

B Details of Evaluation Metrics

	Trig-C		
	P	R	F1
ChatIE	50.5	41.5	45.6

Table 10: Trigger classification results of ChatIE on ACE05 dataset.

RE. We report the standard micro F1 measure and adopt two evaluate metrics: 1) *border* evaluation (BE): an extracted relation triple *(subject, relation, object)* is considered as correct if the whole entity span of both subject and object and relation are all correct. 2) *strict* evaluation (SE): in addition to what is required in the *border* evaluation, the type of both subject and object also must be correct. We use BE on NYT11-HRL because there is no annotation of entity types and use SE on DuIE2.0.

NER. We consider the complete matching and use the micro F1. Only when both the border and the type of the predicted entity and the true entity are the same will we regard it as a correct prediction.

EE. We adopt the different evaluation metrics on the DuEE1.0 dataset and ACE05 dataset. For the DuEE1.0 dataset, F-measure (F1 [6]) is scored according to the word-level matching. For the ACE05 dataset, the predicted argument results are matched with the manually marked argument results at the entity level and evaluated by the micro F1.

C Details of Results

NER. The baseline approach on NER is AdaSeq Bert-CRF on both datasets. We train AdaSeq Bert-CRF in different settings to get the few/full-shot performances in Tab. 1 (Row fs-1/5/10/20/50/100 and full-shot). We also provide some supervised approaches for reference: Noise-robust Co-regularization + LUKE(Zhou and Chen, 2021), BERT-MRC+DSC (Li et al., 2020b), Baseline + BS (Zhu and Li, 2022) and W2NER (Li et al., 2021), shown in Tab. 12. The Sup-SOTA approaches shown in Tab. 1 are Noise-robust Co-regularization + LUKE and BERT-MRC+DSC for conllpp and MSRA, respectively.

RE. The two baseline approaches (*i.e.*, PaddleNLP LIC2021 IE and CasRel) are trained on DuIE2.0 and NYT11-HRL, respectively, for fs-1/5/10/20/50/100 and full-shot. For the results in Tab. 1, the full-shot results on NYT11-HRL and

Dataset	Model	Trig-C			Arg-C		
		P	R	F1	P	R	F1
ACE05	EEQA (Du and Cardie, 2020)	71.1	73.7	72.4	56.8	50.2	53.3
	Text2Event-T5-base (Lu et al., 2021)	67.5	71.2	69.2	46.7	53.4	49.8
	Text2Event-T5-large (Lu et al., 2021)	69.6	74.4	71.9	52.5	55.2	53.8
	DeepStruct (Wang et al., 2022)	-	-	69.2	-	-	63.9
DuEE1.0	GFEE	-	-	-	84.56	83.57	84.06
	ReLiNk	-	-	-	82.12	87.00	84.49
	HIKNLU	-	-	-	86.02	84.41	85.21

Table 11: Result of SOTA supervised approaches for EE.

Dataset	Model	P	R	F1
conllpp	Noise-robust Co-regularization + LUKE(Zhou and Chen, 2021)	-	-	95.88
MSRA	BERT-MRC+DSC (Li et al., 2020b)	-	-	96.7
	Baseline + BS (Zhu and Li, 2022)	-	-	96.3
	W2NER (Li et al., 2021)	-	-	96.1

Table 12: Result of SOTA supervised approaches for NER.

Model	P	R	F1
FCM (Gormley et al., 2015)	43.2	29.4	35.0
MultiR (Hoffmann et al., 2011)	32.8	30.6	31.7
TPLinker (Wang et al., 2020)(exact)	55.43	55.12	55.28
CasRel (Wei et al., 2020b)(exact)	47.88	55.13	51.25
RERE (Xie et al., 2021)(exact)	52.40	58.91	**55.47**
HIKNLU(exact)	82.44	80.68	**81.55**
ReLiNk(exact)	83.16	75.75	79.28

Table 13: Result of SOTA supervised approaches for RE. Note that "exact" means exact match.

Sup-SOTA are reported in the original paper. For other settings, we re-implement the model and report the experimental results. We provide more supervised approaches in Tab. 13 for reference. The top block shows results on NYT11-HRL[16], where the models are the same as Wei et al. (2020b); Xie et al. (2021). The bottom block shows results for the Chinese dataset DuIE2.0, where the models from teams "ReLiNK, HIKNLU" can be found on the official website of AI Studio competition[17]. RERE and HIKNLU are the Sup-SOTA approaches shown in Tab. 1 for NYT11-HRL and DuIE2.0, respectively.

It is worth noting that since NYT11-HRL is obtained by remote supervision, the gold label is not complete and does not cover all relationships (Wei et al., 2020b). We show an example here. For the sentence *He is survived by his wife, Linda, and daughters, Sharon Kofmehl of Charleston, SC, and Sandy Kofmehl of Paris, France, and granddaughter, Emma Kofmehl of Charleston.*, the golden labels only contains one triple: *[(France, /location/location/contains, Paris)]*. However, the predicted output of our model includes more than one triple: *[(France, /location/location/contains, Paris), (SC, /location/location/contains, Charleston), (Sandy Kofmehl, /people/person/place_lived, Paris),...]*. The last several triples should be annotated but were omitted, which significantly affects our model's precision, recall, and F1.

EE. The two baseline approaches (*i.e.*, PaddleNLP LIC2021 EE and Text2Event-T5-base) are for DuEE1.0 dataset and ACE05 dataset, respectively. For the EE results in Tab. 1, only the results of Sup-SOTA are reported in the original paper or technical report. We provide more supervised approaches in Tab. 11. Where the models from teams "GFEE, ReLiNK, HIKNLU" can be found on the official website of AI Studio[18]. The Sup-SOTA approaches shown in Tab. 1 on DuEE1.0 and ACE05

[16] https://paperswithcode.com/sota/relation-extraction-on-nyt11-hrl

[17] https://aistudio.baidu.com/aistudio/competition/detail/46/0/leaderboard

[18] https://aistudio.baidu.com/aistudio/competition/detail/46/0/leaderboard

are EEQA and HIKNLU, respectively. Thus we show them in Tab. 1. In addition, we report the trigger classification results of ChatIE on ACE05 in Tab. 10.