State of relation extraction using LLMs: A report

Vangmay Sachan*[1] and Yanfei Dong †[1]

[1] *National University of Singapore*
[2] *Odyssey 2023/2024*

Abstract

Relation Extraction is one of the three key tasks within the field of information extraction. Relation extraction aims to extract contextual relation between entities from plain natural language texts. Large language models have demonstrated impressive ability to understand the task and generate remarkable results. Because of their ability, numerous works have proposed different techniques and methods to train LLMs in the task of relation extraction.

In this report, we survey some of the advancements in this field. First, we present an overview of techniques and benchmarks used for relation extraction then we categorise them and analyse the most promising methods. We also recognize the fallbacks and possible future directions for the field.

1 Introduction

Information extraction is a crucial domain in natural language processing which converts unstructured data into structured knowledge (eg, entities, relations and event) and serves as a foundational requirement for other downstream tasks. It comprises three key tasks: Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE). In this paper, we are going to focus on the sub-domain of Relation Extraction. The surge of large language models (LLMs) (eg, GPT-4 and Flan T5) has promoted the development of NLP, due to their extraordinary capabilities in text understanding and generation. Because of such advancements, there has been a wave of interest in generative methods to solve relation extraction. Because of their wide language ability, LLMs are able to understand and account for the complexity of human language which makes them more practical in real world applications as compared to other methods. They are also easily able to handle schemas as varying lengths of input without degradation in performance unlike their predecessors (RNNs)

In this report, we provide an exploration of the usage of LLMs in generative RE. We categorise LLM techniques into 3 taxonomies: One-Shot / Few-Shot, Alignment, UIE.

2 Preliminaries of relation extraction

2.1 What is relation extraction

Relation extraction is the task of extracting contextual information from unstructured data. Such as if a text says "John works for apple" then the output should be("John", "works_for", "apple").

Relation Classification

Refers to classifying the relation type between two given entities

Relation Triplet

Refers to identifying the relation type and the corresponding head and tail entity spans

Relation Strict

Refers to giving the correction relation type, the span and the type of head and tail entity

2.2 Document Level + Sentence Level

Relation extraction is being studied upon at two different levels, document level and sentence level. Document level is more challenging because the model needs to look at "evidence" which can be present throughout the entire document no matter how long it is.

2.3 Overview of current benchmarks

There are several benchmarks being used to evaluate relation extraction. Most representative of them are

DocRED

Document-Level Relation Extraction Dataset is a relation extraction dataset constructed from Wikipedia and

*vangmay.sachan16@email.com
†

Dataset	SoTA Model	Score
DocRED	DREEAM	67.530
TACRED	RAG4RE	86.600
NYT	UniRel	93.700
CoNLL04	REBEL	76.650
ACE 2005	PL Marker	73.000

Table 1: Caption

Wikidata. Each document in the dataset is human-annotated with named entity mentions, coreference information, intra- and inter-sentence relations, and supporting evidence. DocRED requires reading multiple sentences in a document to extract entities and infer their relations by synthesising all information of the document. Along with the human-annotated data, the dataset provides large-scale distantly supervised data.

TACRED is a large-scale relation extraction dataset with 106,264 examples built over newswire and web text from the corpus used in the yearly TAC Knowledge Base Population (TAC KBP) challenges. Examples in TACRED cover 41 relation types as used in the TAC KBP challenges (e.g., per:schools_attended and org:members) or are labelled as no_relation if no defined relation is held. These examples are created by combining available human annotations from the TAC KBP challenges and crowdsourcing.

New York Times Annotated Corpus contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007 with article metadata provided by the New York Times Newsroom, the New York Times Indexing Service and the online production staff at nytimes.com. The corpus includes:

- Over 1.8 million articles (excluding wire services articles that appeared during the covered period).
- Over 650,000 article summaries written by library scientists.
- Over 1,500,000 articles manually tagged by library scientists with tags drawn from a normalised indexing vocabulary of people, organisations, locations and topic descriptors.
- Over 275,000 algorithmically-tagged articles that have been hand verified by the online production staff at nytimes.com. As part of the New York Times' indexing procedures, most articles are manually summarised and tagged by a staff of library scientists. This collection contains over 650,000 article-summary pairs which may prove to be useful in the development and evaluation of algorithms for automated document summarization. Also, over 1.5 million documents have at least one tag. Articles are tagged for persons, places, organisations, titles and topics using a controlled vocabulary that is applied consistently across articles. For instance if one article mentions "Bill Clinton" and another refers to "President William Jefferson Clinton", both articles will be tagged with "CLINTON, BILL".

CoNLL04 dataset is a benchmark dataset used for relation extraction tasks. It contains 1,437 sentences, each of which has at least one relation. The sentences are annotated with information about entities and their corresponding relation types.

[Table of representative models to benchmarks]

2.4 Overview of current approaches

2.4.1 Few shot / One-shot

Few Shot approaches can be categorised as a prompt engineering approaches. The main aim is to craft elaborate prompts and also give the model examples and expect the model to follow those examples when producing the output. Some studies have experimented with Chain of thought style examples which use logical thinking in the same manner as humans do and have gained good results from it. Besides that, some approaches also involve fine-tuning the model on those prompts for complex situations which require more elaborate instructions.

2.4.2 Alignment

There is a key issue in Relation extraction, which is that it is rare to find datasets whose design is suitable for relation extraction. Many researchers have hypothesised that because of this performance of LLMs in relation extraction is not as good as their performance in other tasks. This is especially observed in zero shot scenarios. To combat this issue, the approach being taken is to formulate relation extract as other IE tasks which are prevalent in a higher number of datasets.

2.4.3 Universal Information Extraction

As mentioned for Alignment tasks, that very few datasets are built keeping relation extraction in mind, this is not only true for RE but also several other tasks. The problem with the field of information extraction is that despite the fact that these tasks are very "close" in the sense they require the same abilities but the scientific community has came up with different infrastructures like datasets, models, techniques and these tasks have become very disconnected from each other. Universal information extraction or UIE for short is a framework that was first explored in the paper (Lu et al., 2022). Which aims to unify these tasks to ensure we are able to share knowledge from one task to another. Basically, build a framework that can do all tasks within the domain of Information extraction.

3 Techniques of LLMs for Generative RE

In this section we categorise recent methods based on their technique, including Prompt design, Zero Shot learning, Data augmentation, supervised fine tuning and UIE. We also give an analysis of the recent methods by looking at papers and where does their future direction lead to.

3.1 Prompt Design

Prompt engineering is a technique employed where instead of altering model parameters, we try to design creative prompts to exploit LLMs detailed understanding of natural language to guide the behaviour of our model. The craft of prompt design has proven successful in a variety of areas, similarly it plays an effective role in relation extraction.

3.1.1 Chain of thought (CoT)

CoT (Wei et al. 2022b) is a strategy used for LLMs to enhance their performance. It involves adding detailed explanations in the form of a step-by-step reasoning chain which is similar to the human thought process. Ongoing research has explored further into the effectiveness of Chain of thought for Relation Extraction tasks. (Wadhwa et al., 2023) Proposed a distillation technique in which they target RE labels with CoT style explanations elicited from GPT-3 used to fine-tune FLAN-T5 large. A 2-step approach was introduced, leveraging GPT-3 prompts were generated that used chain-of-thought style explanations along with target labels and later the prompts were used to fine-tune Flan-T5 large and as a result, it had a 5 - 10 increase in micro F1 score as compared to SoTA Rebel model (Huguet Cabot and Navigli, 2021).

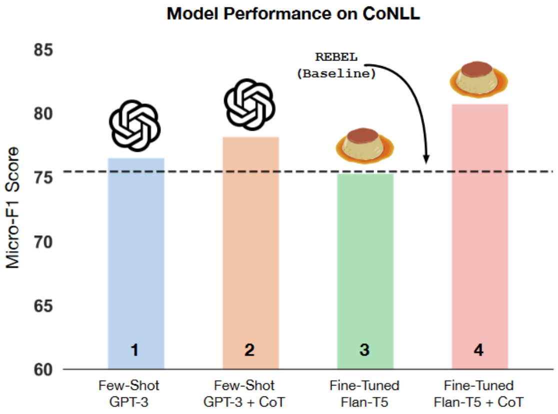


Figure 1: (Wadhwa et al., 2023) Figure 1

The paper highlighted that although Flan-T5 is not as capable in a few shot setting but supervising and fine-tuning it using chain of thought style explanations generated using Chat-GPT yields SoTA results even outperforming GPT-3 which is a much larger model as compared to Flan-T5 large. Moreover, A notable issue was addressed regarding evaluating generated approaches to tasks such as RE using manual human evaluation.

The authors proposed using a few short CoT style prompts, but that led to a performance below SoTA. So the authors proposed fine-tuning Flan T5 largely on a chain of thought style explanations generated by GPT-3 and this led to a major increase across the representative datasets.

- ADE: 9.97 point gain in micro F1 score over the existing fully supervised generative SOTA REBEL.
- CoNLL: 5.42 point gain
- NYT : 3.37 point gain
- Fine tuning Flan T5 with both train labels and CoT explanations produced by GPT-3 yields SOTA performance across RE dataset by 5 - 10 micro F1

Challenges Challenges were faced in automating the evaluation of the model due to the complexity of language. Example if the model is prompted to list all drugs and adverse events it may yield "Aspirin: Stomach ache, chest pain". While another model may return "Side effects of aspirin may include stomach pain and pain in the chest". This challenge was overcome by enlisting human annotators to judge whether the model output conveys the same information as the reference target then compared them to the automated evaluation and categorised the prediction into false-positives, false-negative to compute F1 scores.

3.1.2 Document Level RE

(Li et al., 2023) introduces a sophisticated approach to document level relation extraction (DocRE) through automated annotation leveraging advanced techniques like NLI modules to enhance LLMs. By integrating GPT-3.5 and GPT-4 the framework generates relation triples from text prompts, initially with constraints on entity and relation types that were later relaxed to enhance flexibility and accuracy. These triples are then aligned with predefined annotations in the ReDocRED dataset. Experimental results demonstrate the framework's effectiveness while LLM-generated triples exhibit variability from ground truth, incorporating NLI significance improves alignment, achieving notable recall rates like 5.77 in matching Re-DocRED triples. The resulting dataset, DocGNRE, is enriched with these augmented triples, enhancing the completeness of DocRE datasets. This augmentation is shown to improve the performance of SoTA DocRE datasets in subsequent experiments, underscoring the framework's potential despite challenges like noise in LLM-generated data and the need for robots quality assessment methodologies.

3.1.3 Future directions

Although distillation using Chain of thought style explanations yielded impressive results (Wadhwa et al., 2023), the paper still has some areas where it could be improved. GPT-3 style explanations were not evaluated for correctness More complex RE datasets were excluded, evaluating them on this approach could have led to more insights. The authors did not consider N-ary relations and did not fine-tune GPT3. One interesting observation in the paper was that a small LLMs was able to outperform GPT3 using clever technique, one research direction can be the investigation of other previous tasks looking to find if such observation is true for other smaller LLMs. (Li et al., 2023) Also faced limitations with the number of generated triples being upper bounded.

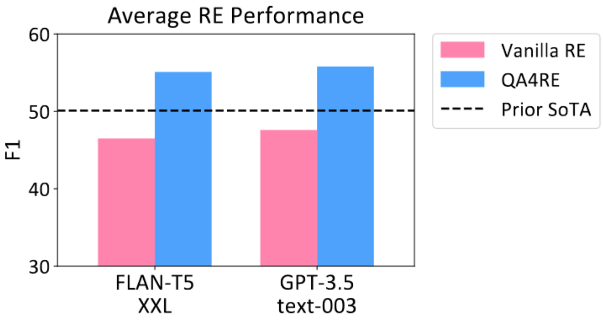


Figure 2: (Zhang et al., 2023) Figure 1 shows the finding that strong Strong instruction-tuned LLMs underperform prior zero-shot RE methods us- ing the standard (vanilla) RE formulation. Meanwhile, QA4RE enables these models to perform better than prior State-of-the-art

3.2 Alignment

Fine tuning LLMs on large scale datasets substantially improves their performance on a wide range of NLP tasks, especially in a zero shot setting. Researchers found that this is not the case for all tasks, especially RE (Zhang et al., 2023). Advanced instruction-tuned LLMs fail to outperform small LMs on relation extraction. Researchers have hypothesised that this occurs because of RE's low incidence in instruction-tuning datasets, making up less than 1%. This has led to methods that formulate RE as other tasks in NLP such that they can benefit from the presence of these tasks in current datasets.

3.2.1 QA4RE

(Zhang et al., 2023) examines the capability of LLMs in identifying the relationship between entities in a sentence. The paper introduces a QA4RE framework to address the low incidence issue which aligns RE with multiple choice question answering (QA), this is because the task appears much more frequently in most instruction tuning datasets. QA4RE brings significant gains over standard RE formulation on validating its effectiveness and the hypothesis regarding low incidence of RE. The framework enabled Davinci003 and FLAN-T5-XXLarge to achieve an average of 8.2% and 8.6% absolute improvement in F1 score.

The paper investigates instruction tuned LLMs on four real world RE datasets and their limited performance on RE might be due to low incidence of RE tasks in instruction tuning datasets. They reformulate RE as multiple choice QA in efforts to exploit its high prevalence in instruction tuning datasets and demonstrate its robustness to diverse prompt designs, they also show how the framework is transferable and consistent across models of different sizes from 80M to 175B.

The experiments required the model to identify the relation between a head entity (E_h) and tail entity (E_t) expressed in S from a set of predefined relation types.

They integrate the head and tail RE entities into relation templates and use them as multiple choice options. The sentence is used as context to the QA system and to ensure fair comparison with previous work, they apply type restrictions to eliminate options for relation types that are not compatible with the entity types of head and tail entities.

For baseline, they used NLI (Sainz et al., 2021) which reformulates RE as a natural language inference task and leverages several LMs fine tuned on MNLI dataset and SuRE (Lu et al., 2022) frames RE as a summarization task and utilises generative LMs such as BART-Large (Lewis et al., 2020) and PEGASUS Large (Zhang et al., 2020), achieving competitive results in few-shot and fully-supervised settings.

By reformulating RE as QA, the framework improves upon the vanilla RE formulation on all the LLMs and most datasets, making them much stronger zero-shot relation extractors.In particular, text-davinci-003 and FLAN-T5 XL and XXL are able to outperform the prior SoTA, NLI DeBERTa, by a large margin. QA4REbrings the largest gain on the best LLM in each series (text-davinci-003 and FLAN-T5 XXL), showing that stronger LLMs may benefit more from our framework. Consistent and substantial improvements can also be observed in other FLAN-T5 models and the full test set. These findings strongly support the hypothesis that aligning underrepresented tasks such as RE unlocks LLMs' ability to solve low frequency tasks.

3.2.2 RAG4RE

A key issue identified by researchers exploring the use of LLMs for RE was the occurrence of hallucinations. LLMs are said to hallucinate when they perceive patterns or objects that are non-existent, leading to nonsensical or inaccurate outputs. (Efeoglu & Paschke, 2024) propose Retrieved-Augmented Generation-based Relation Extraction which aligns RE as a Information retrieval task. The model leverages prominent LLMs including FLANT5, Llama2 and mistral and is evaluated on established benchmarks like TACRED, TACREV, Re-TACRED and semEval RE datasets.

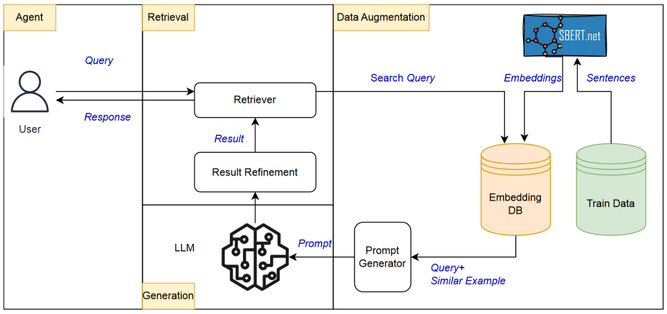


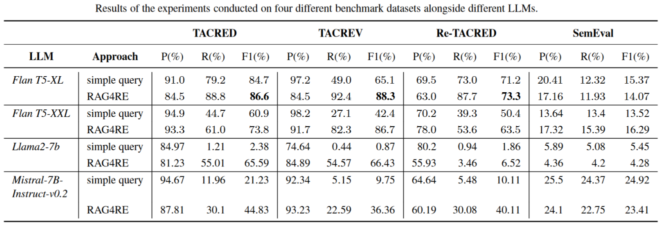
Figure 3: Efeoglu and Paschke (2024) Figure 2, overview of RAG Based relation extraction pipeline

The findings conclude RAG based RE approach has the potential to outperform both simple query known as vanilla LLM prompting and existing best performing RE approach from previous studies and while decoder only LLMs still encounter hallucination issues on these datasets the RAG based RE approach effectively mitigates the problem, especially when compared to the results obtained from the simple query.

The system operates by receiving a user query and a pair of entities that may be related. This query is then passed to a data augmentation module, which utilises a sentence BERT (SBERT) model to compute embeddings for both the query and a large database of training sentences. Using cosine similarity, the module identifies semantically similar sentences from the database and extends the original query with these relevant sentences. These augmented queries, enriched with additional context from the training data, are then passed to a prompt generator. The prompt generator combines the original query with the retrieved sentences to generate a more comprehensive response for the user. This process effectively enhances the user query by leveraging embeddings and similarity scores to incorporate relevant information from a vast dataset, enhancing the relevance and depth of the system's responses. Enhancing the query allows the model to gain more information and reduce the risk of hallucinations occurring, providing robustness to the otherwise unpredictable nature of LLMs.

Results

RAG4RE approach has improved F1 scores on benchmark datasets whose predefined relations are coming from a given sentence tokens (TACREDm TACREVm Re-TACRED) when comparing its results to those of a simple query. RAG4RE has achieved remarkable results when comparing its performance to that of the sim-



Results of the experiments conducted on four different benchmark datasets alongside different LLMs.

		TACRED			TACREV			Re-TACRED			SemEval		
LLM	**Approach**	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Flan T5-XL	simple query	91.0	79.2	84.7	97.2	49.0	65.1	69.5	73.0	71.2	20.41	12.32	15.37
	RAG4RE	84.5	88.8	**86.6**	84.5	92.4	**88.3**	63.0	87.7	**73.3**	17.16	11.93	14.07
Flan T5-XXL	simple query	94.9	44.7	60.9	98.2	27.1	42.4	70.2	39.3	50.4	13.64	13.4	13.52
	RAG4RE	93.3	61.0	73.8	91.7	82.3	86.7	78.0	53.6	63.5	17.32	15.39	16.29
Llama2-7b	simple query	84.97	1.21	2.38	74.64	0.44	0.87	80.2	0.94	1.86	5.89	5.08	5.45
	RAG4RE	81.23	55.01	65.59	84.89	54.57	66.43	55.93	3.46	6.52	4.36	4.2	4.28
Mistral-7B-Instruct-v0.2	simple query	94.67	11.96	21.23	92.34	5.15	9.75	64.64	5.48	10.11	25.5	24.37	24.92
	RAG4RE	87.81	30.1	44.83	93.23	22.59	36.36	60.19	30.08	40.11	24.1	22.75	23.41

Figure 4: Efeoglu and Paschke (2024) Results of the experiments conducted on four different benchmark datasets alongside different LLMs

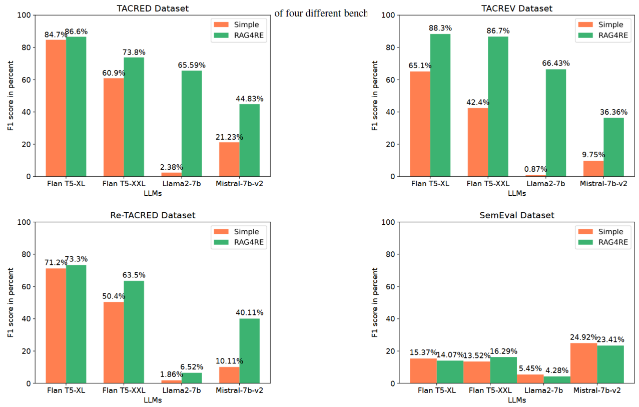


Figure 5: Efeoglu and Paschke (2024) Figure 5

ple query. The highest F1 scores amongst the results at Table 2 have been accomplished 86.6%, 88.3% and 73.3% of F1 scores on the TA-CRED, TARCEVA and Re-TACRED, respectively. These outstanding scores were achieved through the integration of the Flan T5-XL model into the Generation module.

3.2.3 Future directions

Looking at the successful results, future work might involve looking at other under representative tasks and try to align them to pre-existing tasks. We should definitely try experiment by using QA4RE framework using other LLMs such as the OPT series. (Efeoglu & Paschke, 2024) plans to extend their approach (RAG4RE) to real world dynamic learning scenarios and evaluate it using real world datasets.

3.3 Universal Information Extraction

The problem with having different models and different datasets for information extraction tasks is that it reduces knowledge sharing between tasks which are essentially similar and could accelerate progress since one task could benefit from techniques learned for other tasks in information extraction. Universal Information Extraction or UIE is a framework proposed that aims to unify information extraction.

3.3.1 Unified Structure general for information extraction

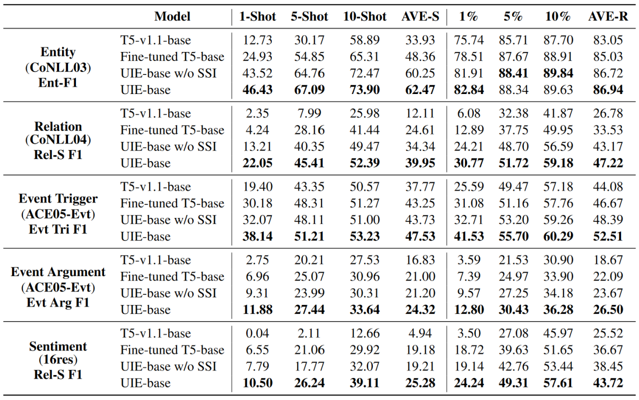
(Lu et al., 2022) Agrees with the notion that IE techniques are specialised which leads to dedicated infrastructure (models, datasets, benchmarks) for each task which hinders the rapid architecture development and knowledge sharing. By recognizing that all IE tasks can be viewed as structure transformations with 2 atomic operations spotting and associating, they develop Structured extraction language (SEL) to model the structure and adaptively generate targeted structures using Structural schema instructor (SSI)

- **Spotting** - Involves locating spans given a semantic type. eg. location in a sentence. "I went to France", ans = France

- **Associating** - Connects spans by assigning them with semantic roles in pre defined schemas. "Steve Jobs" "Apple" = "Work_for". The model is then pretrained using several benchmarks to learn key IE tasks.

SEL proved to be advantageous because it uniformly encodes varying IE structures, hence different IE tasks can be modelled as the same text to structure generation process. Furthermore, SSI constructs a schema based prompt and is used as a prefix during generation, it contains 3 types of tokens (SpotName, AssoName and Special symbols) which then indicate the model of what kind of task it is supposed to accomplish. The model is then pre-trained on several tasks, the datasets are preprocessed such that given a prompt output from SSI, the output expected from the model is in SEL.

Results UIE provides an effective universal architecture for IE, the model achieves SoTA performance on nearly all datasets and tasks, even without pre-training (SEL). The large-scale pre-trained model provides a solid foundation for universal IE. Compared with baselines, the pre-trained model achieves the performance of the state-of-the-art in most datasets and improves 1.42% F1 on average.It proves that SEL is a unified and cross-task transferable structured representation for IE.



	Model	1-Shot	5-Shot	10-Shot	AVE-S	1%	5%	10%	AVE-R
Entity (CoNLL03) Ent-F1	T5-v1.1-base	12.73	30.17	58.89	33.93	75.74	85.71	87.70	83.05
	Fine-tuned T5-base	24.93	54.85	65.31	48.36	78.51	87.67	88.91	85.03
	UIE-base w/o SSI	43.52	64.76	72.47	60.25	81.91	**88.41**	**89.84**	86.72
	UIE-base	**46.43**	**67.09**	**73.90**	**62.47**	**82.84**	88.34	89.63	**86.94**
Relation (CoNLL04) Rel-S F1	T5-v1.1-base	2.35	7.99	25.98	12.11	6.08	32.38	41.87	26.78
	Fine-tuned T5-base	4.24	28.16	41.44	24.61	12.89	37.75	49.95	33.53
	UIE-base w/o SSI	13.21	40.35	49.47	34.34	24.21	48.70	56.59	43.17
	UIE-base	**22.05**	**45.41**	**52.39**	**39.95**	**30.77**	**51.72**	**59.18**	**47.22**
Event Trigger (ACE05-Evt) Evt Tri F1	T5-v1.1-base	19.40	43.35	50.57	37.77	25.59	49.47	57.18	44.08
	Fine-tuned T5-base	30.18	48.31	51.27	43.25	31.08	51.16	57.76	46.67
	UIE-base w/o SSI	32.07	48.11	51.00	43.73	32.71	53.20	59.26	48.39
	UIE-base	**38.14**	**51.21**	**53.23**	**47.53**	**41.53**	**55.70**	**60.29**	**52.51**
Event Argument (ACE05-Evt) Evt Arg F1	T5-v1.1-base	2.75	20.21	27.53	16.83	3.59	21.53	30.90	18.67
	Fine-tuned T5-base	6.96	25.07	30.96	21.00	7.39	24.97	33.90	22.09
	UIE-base w/o SSI	9.31	23.99	30.31	21.20	9.57	27.25	34.18	23.67
	UIE-base	**11.88**	**27.44**	**33.64**	**24.32**	**12.80**	**30.43**	**36.28**	**26.50**
Sentiment (16res) Rel-S F1	T5-v1.1-base	0.04	2.11	12.66	4.94	3.50	27.08	45.97	25.52
	Fine-tuned T5-base	6.55	21.06	29.92	19.18	18.72	39.63	51.65	36.67
	UIE-base w/o SSI	7.79	17.77	32.07	19.21	19.14	42.76	53.44	38.45
	UIE-base	**10.50**	**26.24**	**39.11**	**25.28**	**24.24**	**49.31**	**57.61**	**43.72**

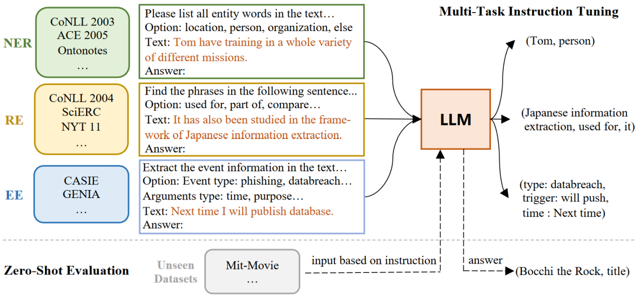
Figure 6: (Lu et al., 2022) Table 3

3.3.2 Instruct UIE

UIE requires separate fine-tuning for different downstream tasks. This leads to the poor performance of UIE in low resources settings or facing new label schema which restricts application of UIE in real scenarios. The approach proposed by (Wang et al., 2023) aims to solve this by reformulating IE tasks as natural language generation problem by introducing a unified IE framework based on instruction tuning, which can uniformly model various information extraction tasks and capture the inter task dependency. The results demonstrated that InstructUIE achieves comparable performance to Bert in a supervised setting. The method outperformed current SoTA and GPT-3.5 in a zero shot scenario.

The IE tasks are formulated as seq2seq tasks and every task instance is formatted with 4 properties. Instruction: Provides a guide on how to extract the relevant information from the input text and produce the desired output structure. Options: Are output labels constraints for a task they represent the set of possible outputs that can be generated by the model. For example: Relation types Text: Input sentence of the task instance. The sequence is then fed into the pre-trained language model along with the task instructions and options, enabling the model to generate the desired output sequence for the task.

NER The model achieves an average F1 score of 85.19% on 20 NER datasets, surpassing Bert's 80.09% where InstructUIE achieved 92.94%. InstructUIE outperforms the Bert model on 17 of NER datasets. Among them the model outperforms Bert by more than 5 points on 8 datasets. The dataset with biggest gap is the broad twitter dataset where the different is of 25 points. Relation Extraction Model reaches an average of 67.98% on the eight datasets of RE tasks, among which the NYT dataset reaches 90.47% F1 score. For the CoNLL2004 dataset, InstructUIE outperforms UIE by more than three points.

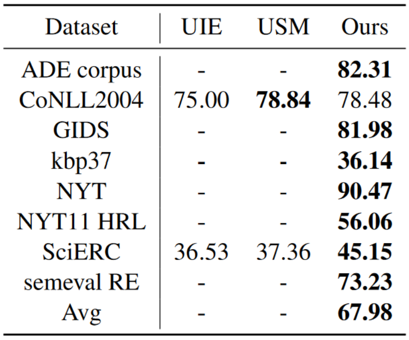


Please list all entity words in the text…
Option: location, person, organization, else
Text: Tom have training in a whole variety of different missions.
Answer:

Find the phrases in the following sentence...
Option: used for, part of, compare…
Text: It has also been studied in the framework of Japanese information extraction.
Answer:

Extract the event information in the text…
Option: Event type: phishing, databreach…
Arguments type: time, purpose…
Text: Next time I will publish database.
Answer:

Figure 7: Wang et al. (2023) Figure 2



Dataset	UIE	USM	Ours
ADE corpus	-	-	**82.31**
CoNLL2004	75.00	**78.84**	78.48
GIDS	-	-	**81.98**
kbp37	-	-	**36.14**
NYT	-	-	**90.47**
NYT11 HRL	-	-	**56.06**
SciERC	36.53	37.36	**45.15**
semeval RE	-	-	**73.23**
Avg	-	-	**67.98**

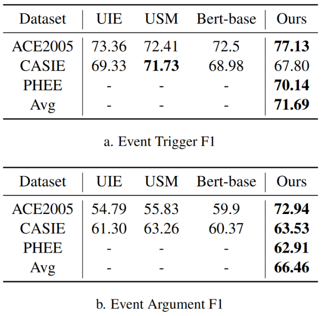
Figure 8: Wang et al. (2023) Table 2

Event extraction The model achieves SoTa on all datasets except for the event trigger F1 metric of the CASIE dataset. On the event trigger F1 metric, the InstructUIE reaches an average of 71.69 on these three dataset. On the event argument F1 metric, instructUIE beats three baseline models to reach SoTA on all three datasets. ACE2005 dataset reaches 72.94%, 18 points higher than the UIE and 17 points higher than USM.

Overall, results demonstrate that InstructUIE achieves state-of-the-art results under supervised and zero settings and solves massive tasks using a single multi-task model.

3.4 Future research direction

Future directions in relation extraction using large language models (LLMs) include several key areas of focus. Improved pre-training strategies, especially domain-specific techniques and the use of large-scale datasets, are crucial for enhancing performance.



Dataset	UIE	USM	Bert-base	Ours
ACE2005	73.36	72.41	72.5	**77.13**
CASIE	69.33	**71.73**	68.98	67.80
PHEE	-	-	-	**70.14**
Avg	-	-	-	**71.69**

a. Event Trigger F1

Dataset	UIE	USM	Bert-base	Ours
ACE2005	54.79	55.83	59.9	**72.94**
CASIE	61.30	63.26	60.37	**63.53**
PHEE	-	-	-	**62.91**
Avg	-	-	-	**66.46**

b. Event Argument F1

Figure 9: Wang et al. (2023) Table 3

The evaluation of more complex datasets and the inclusion of multilingual datasets are essential for the advancement of RE. This could include developing models that can handle diverse and intricate linguistic structures across different languages, thereby enhancing the applicability of RE models in global contexts.

Despite significant progress at the sentence level, document-level RE remains a challenging area. Future research could explore more sophisticated approaches to aggregating evidence across entire documents, possibly incorporating techniques from other NLP tasks such as document summarization and coreference resolution.

Extending RE models to be more KB-aware (knowledge base-aware) can significantly improve their performance. Future work could focus on integrating external knowledge bases into RE models to enhance their ability to accurately extract and classify relations, especially in cases where the context or entities are not well-represented in the training data.

Building upon the concept of Universal Information Extraction (UIE), future research could aim to develop more comprehensive frameworks that unify various IE tasks, including entity linking, event extraction, and coreference resolution. This unification could facilitate better knowledge sharing between tasks and lead to more robust and versatile IE systems.

4 Conclusion

In this paper, we have surveyed recent advancements in Relation Extraction (RE), focusing on the impact of large language models (LLMs) and their role in transforming the field. The exploration of various techniques highlights the significant strides made through innovative approaches such as prompt design, alignment, and Universal Information Extraction (UIE).

Our review emphasizes the pivotal role of LLMs in advancing RE tasks. Techniques like Chain of Thought (CoT) and Universal Information Extraction have demonstrated substantial improvements in performance by leveraging the sophisticated understanding and generation capabilities of these models. CoT, in particular, has shown that fine-tuning models with detailed, step-by-step reasoning can lead to impressive gains in accuracy. On the other hand, approaches like QA4RE and RAG4RE have showcased the potential of aligning RE with more prevalent tasks and integrating retrieval-based strategies to overcome limitations such as hallucination.

The paper also identifies key challenges and future directions. Despite the progress, several areas remain ripe for exploration. For instance, the limited evaluation of more complex datasets and the need for improved methods to handle document-level RE highlight ongoing challenges. Additionally, the unification of various IE tasks through frameworks like UIE presents a promising direction for enhancing knowledge sharing and model efficiency. Future research should focus on addressing these challenges by exploring alternative approaches to fine-tuning, expanding the applicability of frameworks like UIE to new contexts, and enhancing the robustness of models against issues like hallucination. Moreover, extending these methods to real-world, dynamic scenarios and incorporating multilingual datasets could further advance the field. In conclusion, the integration of LLMs into RE has opened new avenues for improving performance and efficiency. The continuous evolution of these models and techniques promises to drive further innovations, making RE more accurate and applicable across diverse domains. The ongoing efforts to refine these methods and address existing challenges will be crucial in advancing the state-of-the-art and achieving more comprehensive and reliable information extraction solutions.

5 What did I learn from odyssey

Doing odyssey for summer 2023 was my first step towards a career of research and a very hands-on experience of how research looks like. I was given weekly tasks which involved reading papers, identifying their key information and presenting it to my supervisor then discussing the next steps in the project. It was a very illuminating experience into the world of research. I would like to thank my supervisor Ms Yanfei Dong as she helped me navigate through the project, from deciding the project to helping me improve my presentation and writing skills by providing valuable feedback at every step and teaching me how to use research tools to my advantage and conduct a literature review.

References

[1] Wadhwa S, Amir S, Wallace B. Revisiting Relation Extraction in the era of Large Language Models. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 15566-89. Available from: https://aclanthology.org/2023.acl-long.868.

[2] Li J, Jia Z, Zheng Z. Semi-automatic Data Enhancement for Document-Level Relation Extraction with Distant Supervision from Large Language Models. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 5495-505. Available from: https://aclanthology.org/2023.emnlp-main.334.

[3] Zhang K, Gutiérrez BJ, Su Y. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors; 2023. Available from: https://arxiv.org/abs/2305.11159.

[4] Efeoglu S, Paschke A. Retrieval-Augmented Generation-based Relation Extraction; 2024. Available from: https://arxiv.org/abs/2404.13397.

[5] Lu Y, Liu Q, Dai D, Xiao X, Lin H, Han X, et al.. Unified Structure Generation for Universal Information Extraction; 2022. Available from: https://arxiv.org/abs/2203.12277.

[6] Wang X, Zhou W, Zu C, Xia H, Chen T, Zhang Y, et al.. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction; 2023. Available from: https://arxiv.org/abs/2304.08085.