From Syntax to Semantics: Evaluating the Impact of Linguistic Structures on LLM-Based Information Extraction

Anushka Swarup[1], Avanti Bhandarkar[1], Ronald Wilson[1], Tianyu Pan[1], Damon L. Woodard[1]

[1]Florida Institute for National Security, University of Florida, Gainesville FL 32611, USA.

Correspondence: aswarup@ufl.edu

Abstract

Large Language Models (LLMs) have brought significant breakthroughs across all areas of Natural Language Processing (NLP), including Information Extraction (IE). However, knowledge gaps remain regarding their effectiveness in extracting entity-relation triplets, i.e. Joint Relation Extraction (JRE). JRE has been a key operation in creating knowledge bases that can be used to enhance Retrieval Augmented Generation (RAG) systems. Prior work highlights low-quality triplets generated by LLMs. Thus, this work investigates the impact of incorporating linguistic structures, such as constituency and dependency trees and semantic role labeling, to enhance the quality of the extracted triplets. The findings suggest that incorporating specific structural information enhances the uniqueness and topical relevance of the triplets, particularly in scenarios where multiple relationships are present[1].

1 Introduction

IE is a crucial NLP task that extracts structured knowledge from unstructured data. Named Entity Recognition (NER) and Relation Extraction (RE) are two essential sub-processes that facilitate IE. They play an integral role in the population of knowledge bases (KBs), where entities serve as nodes and relationships as connecting links. Most recently, NER and RE have been employed to create knowledge graphs for GraphRAG applications (Edge et al., 2024; Han et al., 2024). A popular paradigm, JRE, unifies NER and RE by identifying entities and relationships in a single task from the text sample. JRE garnered significant attention from the NLP community before LLMs emerged. However, research on this topic has become scarce with the advent of LLMs, leaving a gap in understanding their impact on JRE.

The limited studies exploring LLMs for JRE have highlighted issues such as redundancy in extracted triplets and low topical similarity to the target sentence (Swarup et al., 2025). These issues are further exacerbated when there is a possibility of multiple relationships. These problems, coupled with the challenges in evaluation caused by the open-ended nature of the LLMs (Wadhwa et al., 2023), have limited their usage for this task. However, LLMs possess vast knowledge and strong instruction-following capabilities, suggesting their potential to serve as effective joint extractors.

Linguistic structures have been employed as an aid for IE systems through the advancements in NLP. On the one hand, dependency trees (DT) have imparted fine-grained knowledge about the connections between the words in a sentence to neural networks (Tian et al., 2021; Miwa and Bansal, 2016; Chen et al., 2021). On the other hand, Semantic Role Labeling (SRL) has shown close connections with OpenIE triplets (Christensen et al., 2011, 2010). Additionally, constituency trees (CT) have provided a structured representation of input text to the models (Jiang and Diesner, 2019).

This work investigates the potential of using linguistic structures CT, DT, and SRL as additional knowledge to the LLMs to enhance their performance on sentence-level JRE. The goal is to improve their fine-grained semantic understanding from the comparatively shorter context present at the sentence level. Our findings indicate that structural information improves the quality of extracted triplets, especially with smaller models. Additionally, these linguistic structures help LLMs better handle text with multiple relationships, ensuring more accurate and contextually relevant triplets.

2 Methodology

Problem Statement: Given a sentence S and an LLM-based joint relation extractor M, the objec-

[1]Code: https://github.com/anushkasw/StructLLM

Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 36–48

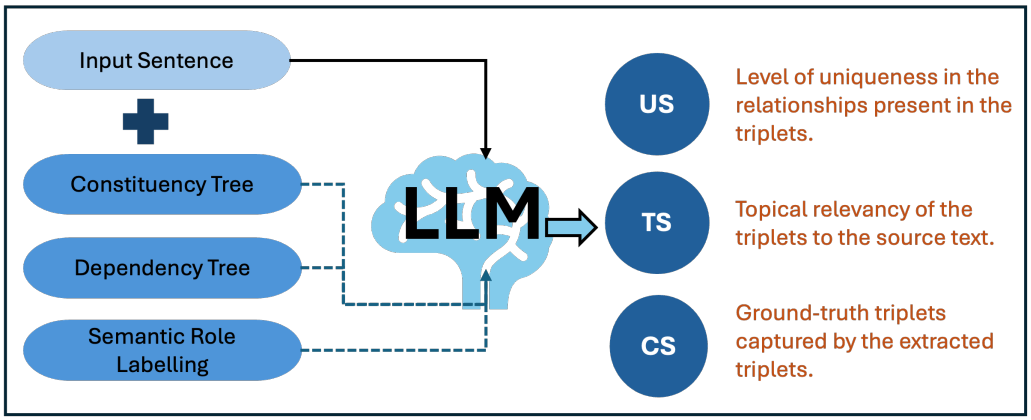

Figure 1: Pipeline to study the influence of linguistic structures on LLMs for JRE.

tive of M is to extract a set of entity-relation triplets of the form $(e1, r, e2)$, where $e1$ and $e2$ are entities, and r represents the relationship between them. Note that information about the entities is not provided apriori. Target relationships may be constrained to a predefined set if required.

2.1 Prompt Engineering

This study investigates the capabilities of LLMs in jointly extracting entities from sentences. A multi-task approach where the prompt instructs the LLMs first to extract all entities from the input sentence and then extract all possible entity and relation triplets was used. This prompting style was used to maximize the LLMs' triplet predictions.

Existing literature shows no significant gain with In-context learning-based (ICL) few-shot strategies for JRE (Li et al., 2023; Swarup et al., 2025). Thus, this work employs zero-shot prompting with two variations. First, when the knowledge of the dataset's relation space is verbalized in the prompt (**rel++**), and second, in an open setting where no such knowledge is provided (**open**). Examples of all prompt types are in Appendix A.3

2.2 Linguistic Structures

Three widely used linguistic structures were selected for this study based on the type of knowledge they provide. First, CTs were incorporated into the LLMs' prompts to assess the impact of syntactic information on JRE. Second, DTs were used to encode both semantic and syntactic relationships, offering valuable insights into word dependencies, even across distant words. Finally, SRL was included due to its strong alignment with the JRE objective, where arguments and verbs can capture the roles of entities and relationships.

2.3 Experimental Setup

Figure 1 depicts the pipeline for this work. This study employs five LLMs as joint relation extractors: OpenChat-3.5 (7B), Meta-Llama-3.1-8B-Instruct (8B), Mistral-Nemo-Instruct-2407 (12B), Gemma-2-9B-IT (9B), and GPT-4o. These models were selected to ensure representation from major LLM families. Multiple parsers were utilized to extract the structural information. Specifically, AllenNLP[2] was used to obtain all three linguistic structures discussed above. Stanza (Qi et al., 2020) was employed to extract CT and DT. As SRL is not supported by stanza, DeepSRL (He et al., 2017) was used as an alternate extractor.

Next, NYT10 (Riedel et al., 2010), TACRED (Zhang et al., 2017), and CrossRE (Bassignana and Plank, 2022) datasets were chosen for this study. Most of these datasets have proven to be challenging for past models. Additionally, the composition of the datasets shows that they have a high percentage of samples consisting of multiple relations, which is a challenging use case. Details regarding the pre-processing steps and statistics of the datasets can be found in Appendix A.1.2.

Subsequently, four experiments were conducted with the configurations: instruction only (**baseline**), instruction+CT (**base+ct**), instruction+DT (**base+dt**), and instruction+SRL (**base+srl**). As the names suggest, these experiments are based on the type of structural information added to the baseline prompt.

Multiple strategies were employed to evaluate the performance of the LLMs for JRE. This work employs traditional metrics (precision, recall, and F1-score) and soft metrics introduced in the Gen-

[2] https://docs.allennlp.org/models/main/

RES benchmark (Jiang et al., 2024). Specifically Uniqueness Score (*US*), Topical Similarity (*TS*), and Completeness Score (*CS*). More implementation details can be found in the source paper and Appendix A.1.4. Finally, the goal of this study was to investigate the performance variations with the *base+dt*, *base+ct* and *base+srl* experiments as compared to the baseline experiments. The Mann–Whitney U test was employed for this purpose to test the statistical significance of the observations ($p\_value < 0.05$).

3 Results

This study investigates the influence of linguistic structures on LLM-based JRE models. As discussed above, both traditional and soft metrics were used to quantify the influence of these elements. Table 2 in the Appendix shows the LLM performances using traditional metrics. The negligible scores attained by the LLMs highlight the impracticality of using these metrics for evaluation. The open-ended nature of the LLM output makes exact matching with ground truth labels almost impossible to perform. Thus, the rest of the study employs soft metrics to assess the quality of the triplets.

Figure 2 depicts the performance of the selected LLMs across datasets and prompting strategies. It can be observed that adding structural information helped enhance the quality of triplets, specifically increasing their uniqueness and topical similarity. However, it had a detrimental effect on the completeness with respect to the ground truths. Fine-grained dataset-specific scores can be found in the Appendix in Table 3.

Structure reduces redundancy in triplets. The findings show that incorporating structural elements such as DT leads to enhancements in triplet extraction across LLMs such as OpenChat, Gemma, and Mistral, as evidenced by the increase in average *US* scores. DT captures crucial syntactical relationships even between distant words in a sentence. This structural information likely enhances the LLMs' comprehension of the text, preventing them from extracting similar relationships. Furthermore, this effect was particularly pronounced for CrossRE, a dataset containing samples from multiple domains, where statistically significant *US* gains were observed.

Structure enhances topical similarity. Incorporating SRL information resulted in more topically relevant triplets across most LLM-dataset combinations, as reflected by the increase in average *TS* scores. Statistically significant gains were observed for Openchat, Llama, Gemma and Mistral across datasets. This suggests that SRL helps LLMs focus more effectively on the language used in the text. SRL outputs closely align with the triplet structure, where arguments often correspond to entities and verbs to relationships. It is likely that LLMs recognize this alignment and leverage SRL to extract triplets. Since SRL outputs are inherently constrained to the sentence context, the LLM's predictions also become more contextually grounded, thereby improving topical relevance.

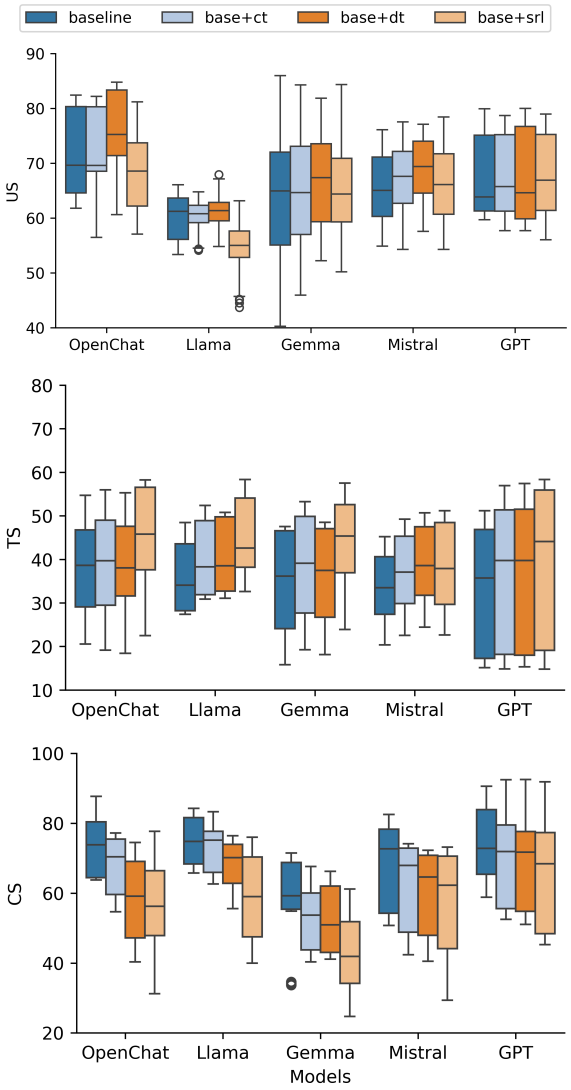

Figure 2: US, TS, CS scores for different LLM-based JRE models across datasets, seeds, and prompting strategies. The x-axis is organized from the LLMs with the fewest parameters to the most parameters. Note that the y-axis range has been adjusted to enhance the visibility of metric variations.

Structure deviates triplets from the ground-truth. The results indicate a linear reduction in *CS* as more advanced structures were incorporated. *CS* measures the comprehensiveness of extracted triplets relative to the ground truth. The observed decline from the *baseline* to *base+srl* suggests that as LLMs integrate more advanced linguistic structures—imparting higher-order and semantic knowledge—their alignment with ground-truth triplets decreases. Previous research has highlighted the limitations of ground truth triplets in most state-of-the-art datasets, suggesting that these triplets are often highly specific and limiting. With the addition of richer semantic information, LLMs tend to generate generalized yet semantically accurate predictions, which may contribute to the decline in *CS*. Additionally, the drop in completeness may also be a byproduct of redundancy reduction. Rather than failing to extract essential information, the model might be filtering out less meaningful triplets.

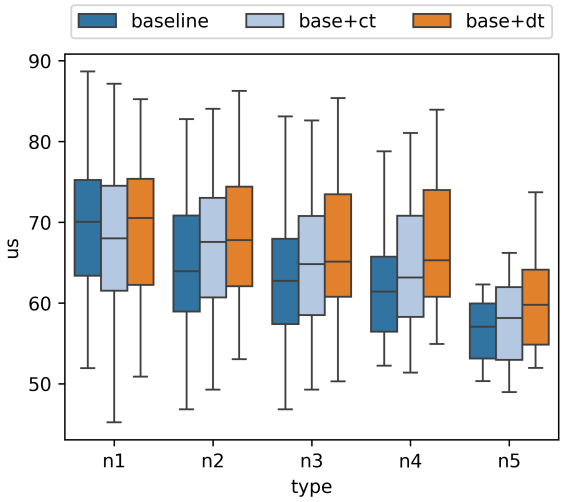

Figure 3: Addition of CT and DT structures prevents uniqueness reduction as the number of relations increases compared to the "baseline" experiments for most categories. Performances are shown across datasets, models, seeds, and prompt types. Note that the y-axis range has been adjusted to enhance the visibility of metric variations.

4 Discussion

The results highlight the redundancy reduction in the extracted triplets, specifically when dependency-based structural information was added to the LLM's prompt. Previous research has highlighted that triplet redundancy increases when the source text contains multiple relationships. To analyze this effect, the *US* scores were examined across sentence categories with varying numbers of ground truth relations. The methodology for this experiment can be found in the Appendix in section A.1.5. Figure 3 indicates that linguistic structures help maintain the uniqueness of extracted triplets compared to baseline LLMs. While the *US* score for *baseline* experiments decreases as the number of relations increases, the *US* scores for LLMs with structural information remain relatively consistent. Additionally, statistically significant gain was observed between the "baseline" and "base+dt" experiments in the *n2-n5* categories. Thus, it can be inferred that linguistic structures help LLMs differentiate between distinct relationships, thereby helping them extract only the most relevant triplets. Some examples of how LLMs can filter out similar meaning triplets by enhancing their semantic reasoning capabilities can be found in Appendix A.2.2.

4.1 Related Works

4.1.1 LLMs for JRE

In recent years, many studies have investigated using LLMs for various IE tasks. Most such studies have focused on the applications of Named Entity Recognition (NRE) (Xie et al., 2023; Kim et al., 2024) and Relation Classification (RC) (Wan et al., 2023; Xu et al., 2023). Very few studies have been conducted for the JRE objective, likely due to the difficulties in evaluation. These studies investigate the potential of using LLMs as zero-shot and few-shot extractors by experimenting with various ICL and prompting strategies (Li et al., 2023; Wadhwa et al., 2023; Swarup et al., 2025). Recently, the GenRES benchmark (Jiang et al., 2024) was proposed to qualitatively evaluate LLM-based JRE extractors.

4.1.2 Structural Modeling for IE

Linguistic structures have been widely used to aid language models for extracting entities and relationships. CTs have been used to provide structured information about the associated text by detailing syntactical knowledge about various grammatical components (Jiang and Diesner, 2019). DTs further enhance this knowledge by highlighting the relationships between words in the sentence, making them a widespread technique in the RC literature (Tian et al., 2021; Miwa and Bansal, 2016; Chen et al., 2021). Finally, SRL has been known to provide semantic understanding of the sentence. They have been employed as a tool in OpenIE systems to

extract entities and relations from the text in an unsupervised setting (Christensen et al., 2011, 2010; Barnickel et al., 2009).

5 Conclusion

This study investigates how incorporating linguistic structures influences LLMs' entity and relation extraction capabilities. The results highlight the improved quality of extracted triplets when structural information is incorporated into the LLMs' prompts. This enhancement in quality was achieved through the reduction of redundant triplets (especially in the presence of multiple relationships) and increased similarity to the source text—both critical in real-world applications. For instance, in a KB construction task from finance data where multiple relationships are common, redundant triplets can create unnecessary paths, reducing the efficiency of the KB. In contrast, topical relevance is critical when extracting knowledge from user-facing systems such as chatbots, making extracting the most relevant entities and relationships essential.

Further, the study highlights a potential drawback: the inclusion of structural information for quality enhancement comes at the cost of misalignment from the ground truth labels, which are oftentimes very restrictive. This finding suggests a trade-off between completeness and uniqueness, which should be carefully considered based on the application's requirements. Finally, there is a need to re-evaluate SOTA datasets in IE, as many contain highly specific and constrained labels. Developing datasets with more generalized label spaces would provide a more comprehensive evaluation framework for IE systems in the LLM era.

Limitations

This study highlights the influence of linguistic structures on LLM performance in JRE, using well-established tools to extract these structures. However, linguistic structure extraction can come with inherent noise, as noted in prior research. However, investigating noise reduction strategies, such as pruning the branches of the trees, was not part of the scope of this paper. This avenue can be explored as future work.

Additionally, we acknowledge that the introduction of the structural elements can introduce noise in the extracted triplets. Thus, an assessment of the factuality of the triplets is imperative. However, existing factualness metrics, including the one present in the GenRES benchmark relies on triplet-level LLM evaluation, which preliminary experiments showed as unreliable and difficult to scale. Despite these challenges, we recognize the importance of this dimensions and aim to find ways to incorporate these metrics in future work.

References

Thorsten Barnickel, Jason Weston, Ronan Collobert, Hans-Werner Mewes, and Volker Stümpflen. 2009. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PloS one*, 4(7):e6393.

Elisa Bassignana and Barbara Plank. 2022. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Guimin Chen, Yuanhe Tian, Yan Song, and Xiang Wan. 2021. Relation extraction with type-aware map memories of word dependencies. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2501–2512, Online. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture*, pages 113–120.

Janara Christensen, Stephen Soderland, Oren Etzioni, et al. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 first international workshop on formalisms and methodology for learning by reading*, pages 52–60.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, et al. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Ming Jiang and Jana Diesner. 2019. A constituency parsing tree based method for relation extraction from abstracts of scholarly publications. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 186–191, Hong Kong. Association for Computational Linguistics.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. GenRES: Rethinking evaluation for generative relation extraction in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2820–2837, Mexico City, Mexico. Association for Computational Linguistics.

Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Jiangming Liu and Yue Zhang. 2017. In-order transition-based constituent parsing. *Transactions of the Association for Computational Linguistics*, 5:413–424.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2019. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

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

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada. Association for Computational Linguistics.

Anushka Swarup, Tianyu Pan, Ronald Wilson, Avanti Bhandarkar, and Damon Woodard. 2025. LLM4RE: A data-centric feasibility study for relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6670–6691, Abu Dhabi, UAE. Association for Computational Linguistics.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021. Dependency-driven relation extraction with attentive graph convolutional networks. 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 4458–4471.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pages 35–45.

A Appendix

A.1 Implementation Details

This section discusses additional details regarding the experimental methodology used for this study.

A.1.1 Model Configurations

Table depicts the detailed model configuration along with the parameters used for this study.

A.1.2 Dataset Details

NYT10, TACRED, and CrossRE datasets were chosen for this study. Since these datasets were created with the RC objective in mind, they had to be converted into a format compatible with JRE. For this, duplicate samples were grouped, and entity and relation triplets were created from the ground truth data provided. Table 1 depicts the test data statistics of the datasets. It can be observed that the grouping facilitated the possibility of multiple triplets associated with the text samples. Finally, according to the trend in the literature, the original test sets were sampled using three random seeds (13, 42, and 100) to lower the cost of LLM processing and provide variability in the experiments. Some additional preprocessing details for the datasets are as follows:

- **NYT10.** The preprocessed version of the dataset (Takanobu et al., 2019) was used for this study.
- **TACRED.** The dataset contains no_relation relationships in the ground truth triplet. It is impractical and redundant for the JRE objective to make the LLM extract triplets where the entities do not have a relationship. Thus, all triplets with the no_relation label were removed at the preprocessing stage.
- **CrossRE.** This dataset was used as-is.

Table 1: Dataset Statistics

Datasets	#n[1]	#r[2]	n1	n2	n3	n4	>n5
CrossRE	913	17	139	134	104	131	405
NYT10	2003	29	1478	307	89	114	15
TACRED	1154	41	841	201	62	25	25

[1] Number of samples [2] Number of relations in the dataset

A.1.3 Parsers

Figure 4 shows how the linguistic structures are incorporated in the LLM's prompt. More details about the parsers used to extract linguistic structures for this study are as follows:

- **AllenNLP.** This study employs a biaffine dependency parser[3] (Dozat and Manning, 2016), constituency parser[4] based on elmo-embeddings (Stern et al., 2017) and BERT-based SRL parser[5] (Shi and Lin, 2019).
- **Stanza[6] (Qi et al., 2020).** This study employs the shift-reduce constituency parser (Liu and Zhang, 2017) and a deep biaffine graph-based dependency parser (Qi et al., 2019).
- **DeepSRL[7] (He et al., 2017).** This study employs an ensemble of deep BiLSTM architecture for SRL.

A.1.4 Evaluation Metrics

As discussed above, TS, CS, and US were used to assess the quality of the extracted JRE triplets. The original methodology (Jiang et al., 2024) was followed to calculate the metrics. Here are some additional details:

- **TS.** This metric was used to quantify the topical similarity of the extracted triplets to the source text. For this, LDA topic modeling was done to extract 150 topics from the test sets of each dataset.
- **US.** This metric was used to calculate the level of uniqueness among the triplets extracted for each test sample. It was calculated by performing similarity matching on the embeddings of the extracted triplets from one another. The triplet embeddings were calculated using OpenAI's *"text-embedding-ada-002"*, and the similarity threshold was set to 0.95.

[3] https://storage.googleapis.com/allennlp-public-models/biaffine-dependency-parser-ptb-2020.04.06.tar.gz

[4] https://storage.googleapis.com/allennlp-public-models/elmo-constituency-parser-2020.02.10.tar.gz

[5] https://storage.googleapis.com/allennlp-public-models/structured-prediction-srl-bert.2020.12.15.tar.gz

[6] https://stanfordnlp.github.io/stanza/index.html

[7] https://github.com/luheng/deep_srl/tree/master?tab=readme-ov-file

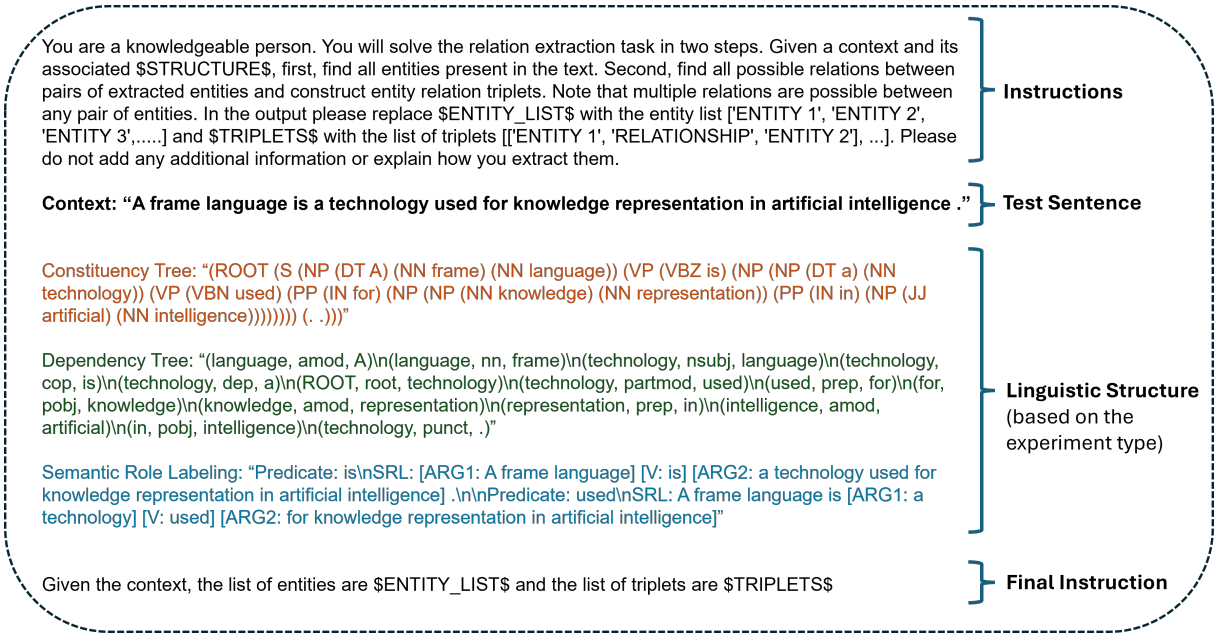

Figure 4: Example of integration of linguistic structures in LLM's prompt.

- **CS.** This metric measures the ground truth triplets covered in the LLMs extraction. It was calculated by performing similarity matching on the extracted triplet embeddings with those of the ground-truth triplets using OpenAI's *"text-embedding-ada-002"* and a similarity threshold of 0.95.

A.1.5 Multiple Relations

To analyze the influence of linguistic structures on samples with multiple relations (discussed in Section 4), we categorized the dataset samples into five groups: n1 (single ground truth relation), n2 (two relations), n3 (three relations), n4 (four relations), and n5 (five or more relations). Table A.1.2 presents the statistics for these categories. Notably, NYT10 contains an insignificant number of samples in the n5 category, while TACRED has very few in both n4 and n5. Consequently, these categories were omitted from calculations for the respective datasets. Additionally, due to the high variability of samples within each category across datasets, we performed 5000 bootstrap experiments by sampling 80, 60, and 100 samples per category for NYT10, TACRED, and CrossRE, respectively, while calculating the metric scores.

A.2 Additional Results

This section showcases results at the dataset and parser-level to provide additional insights.

A.2.1 Dataset-specific Performances

Table 2 depicts dataset-level performances of the LLMs across the four experiments using traditional metrics - precision (P), recall (recall), and F1-score (F1). As mentioned in the main paper, all LLMs attain negligible scores, with GPT as the highest performer. These performances are not a good representation of the LLM extractions. Thus, soft metrics were chosen as the mode of analysis in this study.

Table 3 presents the dataset-level performance of the LLMs across US, TS, and CS dimensions. Overall, the influence of linguistic structures was best observed on the CrossRE dataset. This dataset contains challenging samples from multiple domains such as science, literature, politics, etc., suggesting that linguistic structures can aid in cross-domain understanding of language by LLMs. For the LLMs, smaller models such as OpenChat were most influenced by linguistic knowledge, and the largest model, i.e., GPT, was the least influenced. As the size increases, more and more knowledge is stored in the model's parameters. It is possible that larger models don't require additional assistance as they already contain sufficient knowledge.

A.2.2 Quality Enhancement

Some examples of quality enhancement using linguistic structures are discussed in this section. Table 4 shows two examples where the addition of DT helped reduce the redundancy of the triplets. In the

Table 2: Traditional metric-based evaluation of LLMs for the chosen datasets.

LLM	EXP	NYT10			TACRED			CrossRE		
		P	R	F1	P	R	F1	P	R	F1
GPT	baseline	11.86	16.48	12.67	10.47	14.42	11.19	3.63	4.57	3.78
	base+ct	11.02	15.93	11.98	9.79	14.13	10.63	2.65	3.78	2.87
	base+dep	10.34	15.76	11.42	8.82	13.13	9.66	2.63	3.42	2.78
	base+srl	9.66	14.24	10.57	7.64	11.83	8.55	2.64	3.40	2.78
Gemma	baseline	4.67	6.67	5.01	2.64	3.41	2.67	2.39	2.69	2.35
	base+ct	2.79	4.68	3.12	2.11	3.43	2.29	1.55	2.01	1.60
	base+dep	2.42	3.97	2.69	2.12	3.04	2.23	1.08	1.55	1.13
	base+srl	2.48	4.30	2.85	1.82	2.51	1.91	0.73	0.93	0.73
Llama	baseline	0.64	2.59	0.93	0.10	0.29	0.12	1.12	2.50	1.46
	base+ct	0.49	2.43	0.75	0.10	0.39	0.13	0.73	2.00	0.99
	base+dep	0.34	1.82	0.52	0.06	0.30	0.09	0.50	1.34	0.67
	base+srl	0.38	2.03	0.60	0.03	0.10	0.04	0.60	1.73	0.83
Mistral	baseline	3.24	8.27	4.11	2.24	5.21	2.79	1.97	2.56	2.06
	base+ct	2.30	5.76	2.91	1.98	4.97	2.51	1.96	2.47	2.04
	base+dep	2.22	5.81	2.76	1.48	3.80	1.87	1.45	2.04	1.57
	base+srl	2.41	6.21	3.02	1.77	4.17	2.18	1.51	2.11	1.62
OpenChat	baseline	1.82	6.30	2.50	0.57	2.54	0.86	1.72	2.46	1.89
	base+ct	1.45	5.54	2.07	0.59	2.79	0.92	0.96	1.72	1.14
	base+dep	1.13	4.71	1.67	0.47	2.26	0.74	0.60	0.87	0.65
	base+srl	1.11	4.56	1.63	0.37	1.58	0.57	0.73	1.06	0.79

first case, the addition of the DT helped the LLM extract a diverse set of triplets, which depicted varied relationships. On the other hand, the LLMs without DT could only extract relations of the type *"performed"*. Similarly, in the second example, the standalone LLM can extract relationships only with the entity *"John Mccain"*. However, adding DT helps the LLM extract unique relations such as *"conference calls, participants, bloggers"*, which requires advanced language understanding for extraction. Note that there are cases of erroneous triplets for all experiments, which should be tackled by future work.

Next, Table 5 presents two examples illustrating the topical improvements achieved by incorporating SRL information. These gains can be attributed to the contextual focus that the LLMs attain when structural information is provided. In both examples, the SRL information helps the LLM pay attention to the core topic of each sentence, i.e., the tour and the oil trade, respectively. Without the added structural information, the LLMs tend to focus on general relations, which are more fact-oriented.

A.3 Prompts

Refer to Figures 5-8 for the prompts used for this study.

Table 3: TS, CS, US scores for different LLM-based JRE models for the chosen datasets. * indicates experiments with statistically significant gain as compared to the baseline experiments ($p\_value < 0.05$).

Model	Dataset	Exp	TS	CS	US
OpenChat	NYT10	baseline	32.81	75.57	71
		base+dt	31.74	57.55	72.99
		base+srl	36.94	56.1	68.82
		base+ct	32.47	65.67	68.65
	TACRED	baseline	41.57	69.93	76.12
		base+dt	41.65	58.18	78.52
		base+srl	47.93	53.95	73.64
		base+ct	41.86	67.4	76.05
	CrossRE	baseline	39.29	76.37	64.71
		base+dt	41.11	59.25	75.11*
		base+srl	49.62	61.09	61.69
		base+ct	41.1	70.97	69.24*
Llama	NYT10	baseline	32.94	75.02	61.3
		base+dt	38.2	64.83	62.03
		base+srl	39.04	48.16	51.75
		base+ct	38.14	70.44	61.39
	TACRED	baseline	39	69.68	64.4
		base+dt	41.51	68.98	62.25
		base+srl	47.64	54.52	58.01
		base+ct	41.49	71.04	62.12
	CrossRE	baseline	35.88	80	55
		base+dt	41.11	70.54	59.38*
		base+srl	47.51	66.51	50.38
		base+ct	40.13	78.07	57.21

(continued on next page)

Model	Dataset	Exp	TS	CS	US
Gemma	NYT10	baseline	28.46	57.74	66.38
		base+dt	29.6	52.73	68.76
		base+srl	36.47	42.47	68.54
		base+ct	31.23	51.22	67.68
	TACRED	baseline	35.52	45.37	64.82
		base+dt	36.72	43.65	68.7
		base+srl	45.21	35.47	70.09
		base+ct	39.52	45.31	66.96
	CrossRE	baseline	39.18	69.83	60.36
		base+dt	41	60.9	62.36
		base+srl	49.99*	51.7	57.1
		base+ct	42.77	62.28	59.82
Mistral	NYT10	baseline	28.58	66.39	66.65
		base+dt	33.85	56.48	70.22
		base+srl	31.99	52.82	67.76
		base+ct	32.2	57.79	68.57
	TACRED	baseline	36.21	66.04	72.22
		base+dt	41.02	58.68	74.51
		base+srl	39.25	57.03	72.55
		base+ct	39.21	60.45	73.62
	CrossRE	baseline	35.62	72.96	57.67
		base+dt	40.9	65.35	61.67*
		base+srl	43.03	62.19	58.27
		base+ct	38.95	68.37	58.74
GPT	NYT10	baseline	28.64	77.84	69.11
		base+dt	31.37	72.49	67.83
		base+srl	32.05	69.15	68.11
		base+ct	30.68	72.64	67.94
	TACRED	baseline	34.01	64.4	70.32
		base+dt	36.81	66.42	73.07
		base+srl	38.23	62.55	73.18
		base+ct	36.67	66.62	72.39
	CrossRE	baseline	38.32	79.89	62.07
		base+dt	42.79	71.74	61.22
		base+srl	47.7	68.21	61.9
		base+ct	42.35	73.86	62.11

Table 4: Examples of US enhancement with DT. Only a small set of triples are shown for each prediction.

Sentence	EXP	Model/ Prompt	US	Predicted Triplets
In 2006, along with Sonu Nigam, Sunidhi Chauhan and Shiamak Davar, Ghoshal performed the theme song of 2010 Commonwealth Games at its closing ceremony, as an invitation to everyone to the following Commonwealth Games in Delhi.	baseline	Llama/ rel++	0.16	(ghoshal, performed, 2010 commonwealth games), (ghoshal, performed, commonwealth games in delhi), (ghoshal, performed, everyone), (ghoshal, performed, following commonwealth games), (sonu nigam, performed, 2010 commonwealth games), (sonu nigam, performed, theme song of 2010 commonwealth games)
	base+dt	Llama/ rel++	0.91	(sonu nigam, related to, sunidhi chauhan), (sonu nigam, related to, shiamak davar), (ghoshal, performed, the theme song), (ghoshal, performed, 2010 commonwealth games), (2010 commonwealth games, located in, delhi), (everyone, invited to, the following commonwealth games)
several times this week – while wrapping up a town hall meeting in des moines , or a conference call with bloggers from sioux city , iowa – senator john mccain and his listeners were told that there was time for just one more question.	baseline	Gemma/ open	0.33	(john mccain, is from, iowa), (john mccain, held a meeting in, des moines), (john mccain, had a call with, bloggers from sioux city)
	base+dt	Gemma/ open	0.96	(town hall meeting, location, des moines), (town hall meeting, part of, conference call), (conference call, participants, bloggers), (bloggers, location, sioux city), (time, limited to, one question)

Table 5: Examples of TS enhancement with SRL. Only a small set of triples are shown for each prediction.

Sentence	EXP	Model/ Prompt	TS	Predicted Triplets
A third UK tour for 2017 / 2018 opened at the Curve in Leicester, and also toured to the Birmingham Hippodrome, the Bord Gáis Energy Theatre in Dublin, the Wales Millennium Centre in Cardiff , the Edinburgh Festival Theatre , the Mayflower Theatre in Southampton and the Palace Theatre in Manchester.	baseline	Llama/ open	0.00	(uk, located_in, leicester),(uk, located_in, birmingham hippodrome), (bord gáis energy theatre, located_in, dublin)
	base+srl	Llama/ open	0.88	(a third uk tour for 2017 / 2018, opened at, the curve), (a third uk tour for 2017 / 2018, opened in, leicester), (a third uk tour for 2017 / 2018, toured to, the birmingham hippodrome), (a third uk tour for 2017 / 2018, toured to, the bord gáis energy theatre), (a third uk tour for 2017 / 2018, toured to, dublin)
feud leads to oil cutoff – supplies of russian crude oil headed to european markets came to a halt overnight, in the latest manifestation of rapidly deteriorating relations between russia and belarus.	baseline	Mistral/ rel++	0.00	(russia, person and nationality, belarus), (russia, country and administrative divisions, belarus), (russia, country and capital, moscow)
	base+srl	Mistral/ rel++	0.98	(feud, leads_to, oil cutoff), (russia, produces, crude oil), (russia, exports_to, european markets)

You are a knowledgeable person. You will solve the relation extraction task in two steps. Given a context, first, find all entities present in the text. Second, find all possible relations between pairs of extracted entities and construct entity relation triplets. Note that multiple relations are possible between any pair of entities. In the output please replace $ENTITY_LIST$ with the entity list ['ENTITY 1', 'ENTITY 2', 'ENTITY 3',.....] and $TRIPLETS$ with the list of triplets [['ENTITY 1', 'RELATIONSHIP', 'ENTITY 2'], ...]. Please do not add any additional information or explain how you extract them.

Context: $TEXT$
Given the context, the list of entities are $ENTITY_LIST$ and the list of triplets are $TRIPLETS$

Figure 5: Prompt used for "baseline" experiments in the "open" setting.

You are a knowledgeable person. You will solve the relation extraction task in two steps. Given a context, first, find all entities present in the text. Second, find all possible relations between pairs of extracted entities and construct entity relation triplets. Note that multiple relations are possible between any pair of entities. In the output please replace $ENTITY_LIST$ with the entity list ['ENTITY 1', 'ENTITY 2', 'ENTITY 3',.....] and $TRIPLETS$ with the list of triplets [['ENTITY 1', 'RELATIONSHIP', 'ENTITY 2'], ...]. Please do not add any additional information or explain how you extract them.

Possible Relation Types: $RELATION_SET$

Context: $TEXT$
Given the context, the list of entities are $ENTITY_LIST$ and the list of triplets are $TRIPLETS$

Figure 6: Prompt used for "baseline" experiments in the "rel++" setting.

You are a knowledgeable person. You will solve the relation extraction task in two steps. Given a context and its associated $STRUCTURE$, first, find all entities present in the text. Second, find all possible relations between pairs of extracted entities and construct entity relation triplets. Note that multiple relations are possible between any pair of entities. In the output please replace $ENTITY_LIST$ with the entity list ['ENTITY 1', 'ENTITY 2', 'ENTITY 3',.....] and $TRIPLETS$ with the list of triplets [['ENTITY 1', 'RELATIONSHIP', 'ENTITY 2'], ...]. Please do not add any additional information or explain how you extract them.

Context: $TEXT$
$STRUCTURE$: $STRING$
Given the context, the list of entities are $ENTITY_LIST$ and the list of triplets are $TRIPLETS$

Figure 7: Prompt used for "base+structure" experiments in the "open" setting.

You are a knowledgeable person. You will solve the relation extraction task in two steps. Given a context and its associated $STRUCTURE$, first, find all entities present in the text. Second, find all possible relations between pairs of extracted entities and construct entity relation triplets. Note that multiple relations are possible between any pair of entities. In the output please replace $ENTITY_LIST$ with the entity list ['ENTITY 1', 'ENTITY 2', 'ENTITY 3',.....] and $TRIPLETS$ with the list of triplets [['ENTITY 1', 'RELATIONSHIP', 'ENTITY 2'], ...]. Please do not add any additional information or explain how you extract them.

Possible Relation Types: $RELATION_SET$

Context: $TEXT$
$STRUCTURE$: $STRING$
Given the context, the list of entities are $ENTITY_LIST$ and the list of triplets are $TRIPLETS$

Figure 8: Prompt used for "base+structure" experiments in the "rel++" setting.