



Enhancing Relation Extraction Through Augmented Data: Large Language Models Unleashed

Manzoor Ali¹ (✉) , Muhammad Sohail Nisar¹ , Muhammad Saleem¹ ,
Diego Moussallem¹ , and Axel-Cyrille Ngonga Ngomo¹ 

DICE group, Department of Computer Science, Paderborn University, Germany

manzoor@campus.uni-paderborn.de

msohail@campus.uni-paderborn.de

saleem@informatik.uni-leipzig.de

diego.moussallem@uni-paderborn.de

axel.ngonga@upb.de

<https://www.dice-research.org/>

Abstract. Relation extraction models trained on scarce datasets often exhibit poor performance. The majority of datasets for relation extraction suffer from scarcity, resulting in decreased overall model performance—especially for infrequently encountered relations during training. To alleviate this problem, we present a methodology for augmenting data using large language models by applying in-context learning for relation extraction tasks. Our results reveal that data augmented using different language models can yield distinct effects on relation extraction tasks. Additionally, we compared the performance of the augmented data with other state-of-the-art approaches for data augmentation and conducted a comprehensive analysis of the results. Our findings demonstrate that large language models can produce significantly improved augmented data without the need for fine-tuning, to be utilized in enhancing relation extraction models.

Keywords: Relation extraction · Large language models · Generative AI · In-Context learning · Data augmentation · Prompts

1 Introduction

Relation extraction is a fundamental task within Natural Language Processing (NLP) that finds extensive application across various domains. Its utility extends to information extraction [12], knowledge graph construction and completion [18], question answering [24], and a multitude of other language understanding tasks. Predicting the semantic relationship between named entities in a passage of natural language text, while taking into consideration the context where the entities appear, is the primary objective of relation extraction. For example, *The company Apple Inc. is headquartered in Cupertino, California.* Where two

entities are "Apple Inc." and "Cupertino, California" while the semantic relation is *headquarter*.

Over the past two decades, the field of relation extraction has seen the adoption of various methodologies, each with its own merits and limitations. Among these approaches, supervised methods have consistently demonstrated superior performance [13,2]. However, they come with a significant caveat, as these models rely heavily on the availability of high-quality labeled training data. Acquiring such data is often not only difficult, but also prohibitively expensive [13]. A common challenge encountered in relation extraction is the scarcity of relevant datasets [1]. Many existing datasets used for relation extraction suffer from a shortage of labeled examples, which in turn adversely impacts the overall performance of models. This scarcity is especially detrimental when it comes to relations that are seldom encountered in the training data [11].

To address this issue, data augmentation has emerged as a valuable strategy, not only in the broader realm of machine learning but also specifically within relation extraction. Various data augmentation techniques have been devised, falling into two general categories [4]:

Feature-Based approaches: These methods aim to artificially enhance dataset size by introducing modifications to the linguistic features of the text [10]. Such modifications might involve altering synonyms, adding or removing adjectives, reordering words, or incorporating semantically related terms. However, a primary limitation of these approaches is that they do not generate entirely new sentences; instead, they only augment the existing data, which may not be as effective as other methods.

Language Models based approaches: Another class of approaches involves fine-tuning existing language models to augment the data [9]. While this approach holds promise, it often falls short of task-specific augmentation.

In recent developments, there has been a paradigm shift toward harnessing the generative capabilities of large language models. These models, with their extensive knowledge of language, are being employed to produce augmented data. This research addresses the challenge of data scarcity in relation extraction by introducing an approach to data augmentation that leverages the power of large language models through an in-context learning mechanism to generate augmented data for relation extraction, to improve model performance in scenarios marked by data scarcity and reduce the reliance on high-quality labeled examples.

2 Related Work

Edit-Based Methods apply rule-based changes to original utterances, creating new variations in the data introduced by [22]. *Backtranslation Approaches* [17] rely on translating examples from one language to another and back again. This technique leverages multilingual data to augment the training set. *Fine-Tuned Language Models* such as GPT-2 is used to generate augmented data directly from pre-existing models [8]. This approach has become a cornerstone in data

augmentation for NLP tasks. *Prompt-based Approaches*, such as GPT3Mix by [23], utilize prompts containing lists of possible classes to generate augmented examples and labels. However, challenges may arise when scaling such approaches to set up their approach for multi-class modelling. In the realm of data augmentation, the work of [16] which involves the filtration of unfaithful GPT-generated content, finds alignment with the study by [21]. They introduced a paradigm utilizing GPT-3 for data labeling. In contrast, [16] leveraged GPT-3 not to propose labels for unlabeled samples but to scrutinize and dismiss mislabeled data, specifically within the domain of intent classification.

Data augmentation for RE: Apart from rule-based augmentation techniques, there are some distant supervision-based approaches [25] for relation extraction, which often result in noisy sentences. To the best of our knowledge, there are no augmentation techniques available that utilize open-source LLMs for Relation Extraction (RE) tasks. Wadhwa et al. [20] revisited this area by augmenting labels using the CoT approach on GPT-3, demonstrating overall performance improvement in T5-based fine-tuning. However, they did not compare their label-based augmentation with other augmentation methodologies, nor did they generate new sentences. Josifoski et al. [7] exploited GPT-3.5 to generate independent synthetic data for information extraction pipelines, showcasing improved performance compared to labeled data. However, they did not conduct a comparison between the outputs of different LLMs, particularly those available as open source models. Our method follows the In-context learning approach with more emphasis on understanding sentence complexity.

3 Methodology

Figure 1 presents a comprehensive overview of our methodology employed for data augmentation in relation extraction tasks. Our approach utilizes single-shot in-context learning techniques to augment the data from the original datasets.

1. **Prompt Construction:** We recognize that for RE tasks, it is not required that the data should consist of factually true sentences and may sometimes involve hallucinated content. Therefore, rather than employing a chain of thought approach, we opt for a single-shot method. We construct instruction and combine them with sentences from the original datasets to form prompts. The relation extraction task requires sentences that contain the same type of entities and the similar relation between entities. Therefore, we cannot directly use semantic similarity. For example, the following two sentences:

*The **teacher** graded **papers**; the **chef** cooked **meals**.*

*The **doctor** examined **patients**; the **architect** designed **buildings**.*

are semantically close but do not contain the same relations. Therefore, we need to employ in-context learning to provide further information about the entities types and the relation between the entities. We also offer an instruction-based prompt, wherein we delineate the task and maintain the number of sentences as a variable.

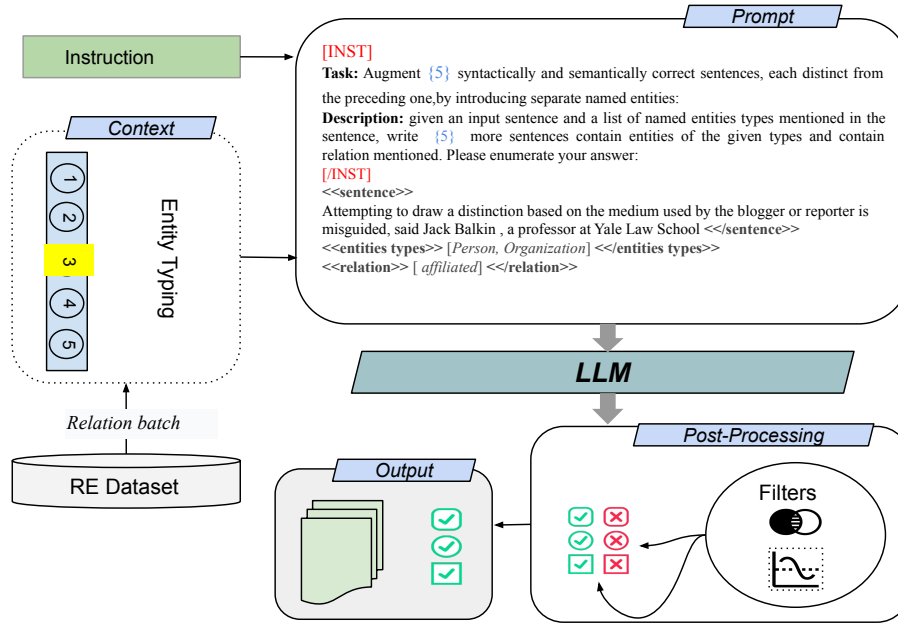


Fig. 1: Overview our methodology, from prompt building to output

2. **Context:** As a context, we provide a sentence from the original dataset to generate augmented data that aligns closely with the training dataset’s context. To mitigate potential biases, we implement an iterative approach for sentence selection. Initially, we categorize the dataset based on relations and employ a *sampling without replacement* method to select a representative sentence for each relation. These chosen sentences are then integrated with associated instruction to form prompts, as illustrated in Figure 1. In cases where the number of available sentences is restricted, such as with the *industry* relation in the NYT-FB dataset, which contains only one sentence, we opt for that sentence during the first iteration. For subsequent iterations, we select sentences from the augmented dataset generated in the initial iteration.

Furthermore, to enrich the context, we leverage a Named Entity Recognition (NER) model¹ to identify the types of entities present in the sentence. This information is included as contextual cues. Additionally, we retain the original relation from the dataset in the prompt to facilitate comprehensive in-context learning.

3. **Complexity and Similarity Filters:** We applied complexity and similarity filters, these filters rely on cosine similarity and Flesch-Kincaid [5] grade complexity to either accept or reject the augmented sentences. The sentences

¹ <https://spacy.io/models>

are augmented in batches, where we calculate the Flesch-Kincaid grade and cosine similarity scores for the entire batch. The Flesch-Kincaid Grade Level Readability Test assesses a text’s level of complexity of comprehension. It gives an approximate idea of the grade level in American schools needed to understand the text. It takes into account the word count as well as the number of syllables in each sentence. The number of years of schooling required for an individual to comprehend the text is indicated by the resulting grade level². We then compute the average scores and compare them to those of the original sentences. If the scores fall within a defined threshold, the sentence is retained. This approach ensures that (see Table 4) the complexity and similarity of the augmented sentences closely align with those of the original dataset.

4 Experimental Setup

Evaluating augmented data for relation extraction is effectively accomplished using a machine learning model. In our experiments, we utilize an RE model based on BERT [3] to demonstrate the quality of the augmented data. We purposely avoided using complex models because simple models benefit more directly from large amounts of data. In contrast, complex models can leverage additional relevant information. Our goal is to demonstrate the advantage of using more data.

Datasets Description:

FewRel Dataset: [6] holds a pivotal role in our experimental design due to its unique characteristics. Notably, FewRel is the sole balanced dataset available for relation extraction, containing 64 training relations. Furthermore, for each of these 64 training relations, the dataset provides a balance of 700 sentences.

NYT-FB Dataset: [15] The selection of the NYT-FB dataset is guided by its distribution of sentences across different relations noisiness, to test our approach’s capability in challenging conditions and its widespread use in RE research. This dataset encompasses a total of 24 relations. Significantly, 11 out of these 24 relations exhibit a limited number of sentences.

Selected Model and Augmented Data In our experimental setup, we employed a foundational transformer-based model that utilizes BERT encoding. Throughout all our experiments, we maintained a consistent set of hyperparameters, ensuring a fair and reliable evaluation. We kept the number of epochs 20 while employing early stopping techniques to mitigate the risk of overfitting. To ensure unbiased evaluation, we exclusively utilized the original test dataset for testing wherever possible in experiments. This approach minimizes potential biases that may arise from the introduction of augmented data. A summary of the hyperparameter configurations is provided on our GitHub³ repository.

We conducted a series of experiments utilizing two open-source distinct Large Language Models (LLMs) for data augmentation: *Llama 34B* [19], and *Falcon*

² For instance, an eighth-grader should be able to comprehend the material if the Flesch-Kincaid Grade Level is 8.0.

³ <https://github.com/dice-group/augmentation-LLM->

40B [14]. Since our primary focus revolves around data augmentation using LLMs, we exclusively generate augmented data using two LLMs capable of addressing our research question. However, our approach remains adaptable to incorporate any state-of-the-art or future LLMs. Additionally, we employed the Natural Language Toolkit (NLTK [10]) to generate an equivalent volume of augmented data.

We kept the following data augmentation strategies: We generated 50 augmented sentences for each relation present in both FewRel and NYT-FB datasets. For relations in the NYT-FB dataset with fewer than 100 sentences, we increased the dataset size by generating additional sentences, ensuring a minimum threshold was met. We report micro-averaged scores for equal augmentation and macro score for augmentation where only sentences increased for limited relations.

5 Results and Discussion

We perform comprehensive experiments to answer the following questions:

RQ₁: Is data augmentation with Large Language Models (LLMs) more effective for relation extraction tasks compared to rule-based approaches? **RQ₂:** Do different LLMs yield similar augmented data, or are there notable differences in the augmented data they produce? **RQ₃:** What is the quantitative impact of LLM-based augmented data on the performance scores of relation extraction models? **RQ₄:** How closely does the augmented data generated by LLMs resemble the original dataset?

FewRel Dataset: Table 1 provides a detailed overview of the Precision (P), Recall (R), and F1-score (F1) metrics for the FewRel dataset across various augmentation scenarios, considering both with and without entity information. The analysis shows the impact of augmenting sentences on the overall performance of the relation extraction model.

As the number of augmented sentences increases, a perceptible improvement is observed in the overall scores for language model-based augmentation. Intriguingly, with only a modest 30% increase, the rule-based approach outperforms language model-based augmentation. This outcome is likely influenced by the inherent complexity of sentences generated by large language models. In scenarios with limited augmentation, the increased complexity tends to enhance the model’s overall performance. However, as the number of sentences augments further, the overall score for rule-based augmentation experiences a decline, while language model-based augmentation continues to effectively enhance the overall score. This observation confirms **RQ₁:** that LLM-based augmentation can surpass rule-based augmentation.

Notably, augmentation using Llama demonstrates superior results compared to Falcon for this specific dataset. The nuanced differences between Llama and Falcon warrant further investigation, as their effectiveness can be dataset-dependent. Exploring the specific characteristics of the FewRel dataset that favor Llama over Falcon could provide valuable insights into the strengths and weaknesses of dif-

Table 1: Precision, Recall, and F1-score for the FewRel dataset at various percentage increases (100% increase mean all the augmented data is used), both with and without entity information. R stands for rule based approach.

Data	With Entities			Without Entities		
	P	R	F1	P	R	F1
FewRel	89.5	89.3	89.3	67.0	66.7	66.4
FewRel(R)	89.4	89.2	89.2	67.7	67.5	67.3
FewRel (Llama) (30%)	88.5	88.4	88.2	66.8	66.7	66.3
FewRel (Falcon)	88.6	88.1	88.0	66.6	0.66	66.0
FewRel(R)	89.6	89.3	89.4	66.3	66.2	65.8
FewRel (Llama) (70%)	90.6	90.2	90.4	68.5	68.0	68.5
FewRel (Falcon)	89.9	89.8	89.8	68.0	0.67	67.3
FewRel(R)	89.4	89.3	89.3	68.9	69.0	68.9
FewRel (Llama) (100%)	91.3	90.0	91.1	70.5	69.5	70.0
FewRel (Falcon)	90.8	90.6	90.7	69.0	69.3	69.14

ferent LLMs based augmentation. This observation answers **RQ₂**: that different LLMs generate different augmentation of sentences.

NYT-FB Dataset: Table 2 provides a detailed examination of the NYT-FB dataset. The scores presented in the table are derived from the test set, which combines augmented sentences with the original training set. It is noteworthy that the original test set lacked sentences for three relations (/people/person/ethnicity, /people/person/profession, /business/company/industry), necessitating the integration of augmented sentences for a comprehensive evaluation. Additionally, we acknowledge that the augmented sentences have not undergone post-augmentation cleaning.

In the case of equal augmentation, both Llama and Falcon demonstrate a significant improvement in performance, surpassing the baseline Precision, Recall, and F1-score. Falcon, in particular, achieves remarkable results with a Precision of 97.6%, Recall of 92.4%, and an outstanding F1-score of 94.2%. These insights show the quantitative impact of data augmentation on the relation extraction task and provide an answer to our third question, **RQ₃**:

Augmentation techniques maintain their superiority even when applied exclusively to relations with fewer than 100 sentences, outperforming baseline results. Falcon stands out with a notable F1-score of 95.7%, underscoring its efficacy in scenarios with limited data. The outstanding performance of Falcon prompts further investigation, where we explore models with 25 sentences per relation for a total of eight relations (200 sentences in total). The selection of these eight relations is based on the top four with the most sentences and the bottom four with the least. The results, presented in Table 3, reveal a consistent improvement for all eight relations compared to the baseline. Notably, Falcon-based augmentation significantly enhances the top four relations, while Llama outperforms Falcon in the bottom four.

Table 2: Precision, Recall, and F1-score, under two scenarios: one with equal data augmentation across all relations, and the other with augmentation limited to relations containing fewer than 100 sentences.

Data	Equi Aug.			#Sent < 100 Aug.		
	P	R	F1	P	R	F1
NYT-FB	60.2	56.7	57.2	-	-	-
NYT-FB (Llama)	79.4	68.7	70.3	80.2	68.5	69.6
NYT-FB (Falcon)	97.6	92.4	94.2	96.0	95.8	95.7

Table 3: Precision, Recall, and F1-score for the eight selected relations in models trained on augmented data the NYT-FB dataset.

Relations	#Sent	Llama			Falcon		
		P	R	F1	P	R	F1
../contains	30240	100	100	100	100	100	100
../nationality	5219	99.4	99.7	99.3	100	100	100
../place_lived	5024	99.4	99.1	99.2	99.5	99.2	99.4
../company	3971	94.4	91.7	93.0	98.8	98.2	98.5
../ethnicity	9	51.4	45.7	48.4	21.4	17.5	19.3
../people	9	49.6	42.4	45.7	19.0	16.8	17.8
../profession	2	39.0	27.7	32.4	9.2	7.5	8.3
../industry	1	35.3	29.7	32.3	8.0	8.0	8.0

An interesting observation is that Falcon produces precisely 100 sentences for each relation, contributing to its exceptional performance. On the other hand, Llama exhibits variations in sentence output, ranging from 41 to 92 sentences per relation. This discrepancy leads to underperformance for the top four relations, where a consistent number of sentences is crucial. However, the varied sentence output of Llama, closely resembling natural language, contributes to its out-performance in relations with fewer sentences.

To address **RQ₄**, we investigate the complexity of the generated sentences using different parameters and assess their semantic coherence with the original dataset sentences. For complexity evaluation, we utilized the Flesch-Kincaid grade to gauge the readability of sentences. Additionally, we determined the average number of entities per sentence, a crucial factor in assessing linguistic intricacy.

Complexity The results of our analysis are detailed in Table 4. Notably, the rule-based approach demonstrated a minimal deviation in the average token count per sentence when compared to the original dataset. This observation aligns with the inherent characteristics of rule-based augmentation, where transformations involve singular word substitutions or positional changes, resulting in sentences that closely resemble their source.

Table 4: Complexity in terms of Avg. number of tokens, entities per sentence and Flesch-Kincaid grade score. Cos sim for semantic coherence

Aug. Tech	Complexity			Cos sim
	Avg #Tokens	Avg. FKg	Avg #entities	
FewRel	25.0	10.34	4.5	0.15
Rule	24.9	10.68	4.10	0.15
Falcon	18.8	11.63	2.4	0.37
Llama	32.8	13.36	3.9	0.18
NYT-FB	37.8	12.89	3.3	0.19
Rule	39.8	14.18	4.2	0.24
Falcon	17.1	11.13	2.2	0.39
Llama	36.2	12.91	3.6	0.16

Semantic Coherence In our exploration of augmented data, we delved into assessing semantic similarity, aiming to uncover the nuances of how well-augmented sentences align with the original dataset and among different augmentation techniques. To quantify this, we employed cosine similarity, a measure for gauging the degree of semantic resemblance. Notably, Llama-based augmentation exhibits a higher degree of semantic similarity to the original dataset when compared to Falcon-based augmentation, which produces sentences that closely resemble each other.

6 Conclusion

In this study, we harnessed the potential of LLMs for data augmentation in the field of relation extraction. We have demonstrated that data augmented with large language models outperforms data augmented using rule-based approaches. We remain committed to addressing the various challenges associated with the utilization of large language models for data augmentation.

Limitations and Ethical Consideration: It is important to note that our approach may not be directly applicable to tasks requiring augmented data with true facts, as we do not prioritize the generation of hallucinated information in the context of relation extraction. On the ethical side, LLM-generated data may exhibit social biases or employ toxic language. Therefore, the augmented data should be used with caution, and proper procedures such as human involvement or debiasing LLMs.

Acknowledgments

This work has been supported by the BMBF-funded project COLIDE (01I521005D) and KIAM (02L19C115), the European Union’s Horizon Europe research and innovation programme ENEXA (101070305), MWIDE NRW funded project Climate bOWL (005-2111-0020), DFG funded project SFB-TRR 318 (438445824) and the University of Malakand Pakistan.

References

1. Ali, M., Saleem, M., Moussallem, D., Sherif, M.A., Ngonga Ngomo, A.C.: Reld: A knowledge graph of relation extraction datasets. In: Pesquita, C., Jimenez-Ruiz, E., McCusker, J., Faria, D., Dragoni, M., Dimou, A., Troncy, R., Hertling, S. (eds.) *The Semantic Web*. pp. 337–353. Springer Nature Switzerland, Cham (2023)
2. Ali, M., Saleem, M., Ngomo, A.C.N.: Unsupervised relation extraction using sentence encoding. In: *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers* 18. pp. 136–140. Springer (2021)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
4. Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. pp. 968–988. Association for Computational Linguistics, Online (Aug 2021). <https://doi.org/10.18653/v1/2021.findings-acl.84>
5. Flesch, R.: Flesch-kincaid readability test. Retrieved October 26(3), 2007 (2007)
6. Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., Sun, M.: Fewrel:a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: *EMNLP* (2018)
7. Josifoski, M., Sakota, M., Peyrard, M., West, R.: Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pp. 1555–1574. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.96>
8. Lee, K., Guu, K., He, L., Dozat, T., Chung, H.W.: Neural data augmentation via example extrapolation (2021)
9. Li, H., Dong, Q., Tang, Z., Wang, C., Zhang, X., Huang, H., Huang, S., Huang, X., Huang, Z., Zhang, D., Gu, Y., Cheng, X., Wang, X., Chen, S.Q., Dong, L., Lu, W., Sui, Z., Wang, B., Lam, W., Wei, F.: Synthetic data (almost) from scratch: Generalized instruction tuning for language models (2024)
10. Loper, E., Bird, S.: Nltk: the natural language toolkit. *arXiv preprint cs/0205028* (2002)
11. Nadgeri, A., Bastos, A., Singh, K., Mulang[?], I.O., Hoffart, J., Shekarpour, S., Saraswat, V.: KGPool: Dynamic knowledge graph context selection for relation extraction. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. pp. 535–548. Association for Computational Linguistics, Online (Aug 2021). <https://doi.org/10.18653/v1/2021.findings-acl.48>
12. Ning, Q., Feng, Z., Roth, D.: A structured learning approach to temporal relation extraction. *arXiv preprint arXiv:1906.04943* (2019)
13. Pawar, S., Palshikar, G.K., Bhattacharyya, P.: Relation extraction: A survey. *arXiv preprint arXiv:1712.05191* (2017)
14. Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only (2023)
15. Riedel, S., Yao, L., McCallum, A., Marlin, B.M.: Relation extraction with matrix factorization and universal schemas. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 74–84 (2013)

16. Sahu, G., Rodriguez, P., Laradji, I., Atighehchian, P., Vazquez, D., Bahdanau, D.: Data augmentation for intent classification with off-the-shelf large language models. In: Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 47–57. Association for Computational Linguistics, Dublin, Ireland (May 2022)
17. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86–96. Association for Computational Linguistics, Berlin, Germany (Aug 2016). <https://doi.org/10.18653/v1/P16-1009>
18. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 455–465 (2012)
19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
20. Wadhwa, S., Amir, S., Wallace, B.: Revisiting relation extraction in the era of large language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15566–15589. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.acl-long.868>
21. Wang, S., Liu, Y., Xu, Y., Zhu, C., Zeng, M.: Want to reduce labeling cost? gpt-3 can help (2021)
22. Wei, J., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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). pp. 6382–6388. Association for Computational Linguistics, Hong Kong, China (Nov 2019). <https://doi.org/10.18653/v1/D19-1670>
23. Yoo, K.M., Park, D., Kang, J., Lee, S.W., Park, W.: GPT3Mix: Leveraging large-scale language models for text augmentation. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2225–2239. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.findings-emnlp.192>
24. Yu, M., Yin, W., Hasan, K.S., Santos, C.d., Xiang, B., Zhou, B.: Improved neural relation detection for knowledge base question answering. arXiv preprint arXiv:1704.06194 (2017)
25. Zhu, T., Wang, H., Yu, J., Zhou, X., Chen, W., Zhang, W., Zhang, M.: Towards accurate and consistent evaluation: A dataset for distantly-supervised relation extraction. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 6436–6447. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). <https://doi.org/10.18653/v1/2020.coling-main.566>