Towards syntax-aware pretraining and prompt engineering for knowledge retrieval from large language models

Stefan Dietze[1,2], Hajira Jabeen[1], Laura Kallmeyer[2] and Stephan Linzbach[1]

[1]GESIS - Leibniz Institute for the Social Sciences

[2]Heinrich-Heine-University Düsseldorf, Germany

Abstract

The ability to access relational knowledge from LLM parameters, known as relational knowledge retrieval (rKR), is considered a critical factor in their capacity to comprehend and interpret natural language. However, the role of syntax in this context has not been adequately explored. In this position paper, we hypothesize a close link between the accessibility of relational knowledge and syntax.

We discuss related works and lay out a research agenda focused on rKR from self-supervised LLMs without or with minimal fine-tuning and aiming at understanding the impact of syntax on rKR. This involves examining biases, factors affecting result reliability and robustness, and analyzing the effect of syntactic features in training corpora on rKR. We argue that rKR can be improved through syntax-aware pretraining and prompt engineering, and propose a dedicated research agenda geared toward exploring the impact of syntax on knowledge retrieval.

1. Introduction

Relational knowledge captures the relations between entities and concepts and is crucial for a wide range of tasks. Traditionally, retrieval and reasoning of relational knowledge have both relied on symbolic knowledge bases [1], that often are constructed using supervised extraction techniques applied to unstructured corpora, e.g. web archives [2, 3].

On the other hand, large language models (LLMs) such as BERT [4], GPT-2 [5], and GPT-3 [6] revolutionized NLP research due to their self-supervised training paradigms and their transferability across various downstream tasks. Recently, LLMs have also been investigated for their ability to directly retrieve relational knowledge [7] from their parameters, e.g. through question answering, prompting through the use of cloze-style questions [8, 9] or statement scoring [10]. In this context, the ability of LLMs to retrieve, infer, and generalize relational knowledge is seen as a crucial indicator of their capacity to understand and interpret natural language. Even though a range of terms is used in that context, e.g. fact or knowledge retrieval as well as knowledge inference, we refer to the task of accessing relational knowledge from LLM parameters as relational knowledge retrieval (rKR).

KBC-LM'23: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2023

stefan.dietze@gesis.org (S. Dietze); hajira.jabeen@gesis.org (H. Jabeen); kallmeyer@phil.uni-duesseldorf.de (L. Kallmeyer); stephan.linzbach@gesis.org (S. Linzbach)

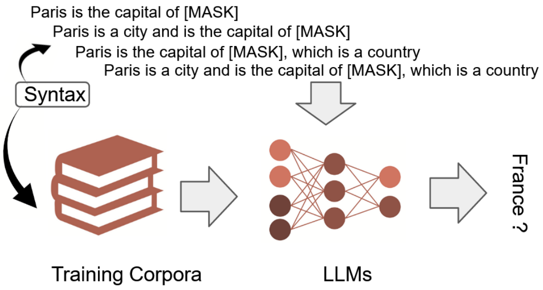
Paris is the capital of [MASK]
Paris is a city and is the capital of [MASK]
Paris is the capital of [MASK], which is a country
Paris is a city and is the capital of [MASK], which is a country
Syntax
Training Corpora
LLMs
France ?

Figure 1: Link between syntax of pretraining and rKR

In addition to learning statistical patterns and relationships among words, LLMs implicitly learn syntactic structure (see, e.g., [11, 12, 13]). While prior work has established that LLMs are capable of retrieving knowledge to some extent, the impact of syntax in that context is under-explored, despite the fact that a link between relational knowledge and syntactic information has been hypothesized [14]. Building on these observations, we assume that the accessibility of relational knowledge and the syntactic structure of pretraining data and prompts are closely linked (see Figure 1). Furthermore, we hypothesize that an increased awareness of syntactic structures increases the LLM's ability to retrieve relational knowledge. For transversal relations, such as *isCreatorOf*, we argue that they correlate with specific syntactic dependency paths. Fig. 2 for instance shows the dependency trees[1] of three different but systematically related syntactic and semantic structures, that each contains the relational knowledge that should enable an LLM to answer the cloze-style prompt (*Orwell*, isCreatorOf, *X*), with "1984". In all three cases, we have specific dependency paths from the relevant nouns ('Orwell' and '1984') to the inflected form of 'write'.

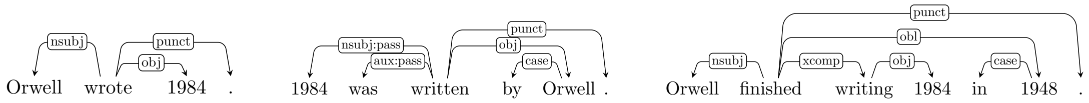

Figure 2: Dependency trees of three distinct sentences containing the same relational knowledge.

For hierarchical relations (e.g., a *president* is a *person*) we hypothesize that whenever an instance of the more general concept can fill a certain syntactic argument slot, this is also possible for the more specific concept, suggesting the utility of ontological knowledge in rKR.

These observations motivate our main position that the potential of LLMs for retrieving and inferring relational knowledge can be significantly increased when considering implicit or explicit information about syntactic structure. Note that, different from various efforts that involve a significant amount of supervised fine-tuning and reinforcement learning from human feedback (RLHF), e.g. InstructGPT [15], we are specifically interested in the generalizations and rKR capacities of self-supervised LLMs without or with minimal fine-tuning while exploiting syntactic structure that has been learned in a supervised way (e.g., via some dependency parser).

With this paper, we motivate and describe a research agenda aimed at investigating the role of syntax for knowledge retrieval from LLMs.

2. Related Work

Benchmarks and baselines for knowledge retrieval from LLMs. LAMA is the first benchmark dataset introduced to evaluate knowledge retrieval in LLMs [16]. Related works show

[1] Obtained via https://corenlp.run/, 03.08.2023

that knowledge retrieval through prompts is inconsistent with regard to paraphrasing [17, 18], with some types of information guiding LLMs towards more correct answers [19, 20, 21], while other types are harmful to their performance [22, 23]. LLMs struggle to retrieve knowledge from low-frequency phenomena [24] and [25] argue that LLMs fail to express large varieties of knowledge when prompted for it in a zero-shot manner. Zhong et al. [26] propose that the models' accuracy may be from memorizing training data, not actually inferring knowledge. Similar to LAMA, experiments on a more recent probe (KAMEL) [27] confirm that LLMs are still far from the knowledge access capabilities of symbolic knowledge bases. The Knowledge Memorization, Identification, and Reasoning test KMIR [28] reveals that while LLMs struggle to robustly recall facts, their capacity to retain information is determined more by the number of parameters than the training methods, and while model compression can help preserve the memorization performance, it reduces the ability to identify and reason about the information in LLMs from transformer-based language models, Linzbach et al. [29] also presents similar findings. LLMs are known to struggle with more complex reasoning tasks [30]. Branco et al. [31] explore the generalisability of common-sense reasoning capabilities and the impact of shortcuts in training data.

The role of syntax in LLMs. LLMs like BERT implicitly learn syntactic information ([13, 32, 33, 12]). Structural information has been shown to aid a range of downstream tasks, such as in our own works leveraging GCN- and attention-based models informed by local and global structure information for sentiment analysis [34] or recommender systems [35], and the work in [36, 37], where we induced event types and semantic roles starting from dependency syntax. Even though LLMs already capture syntactic information to a certain extent (see above), additionally leveraging syntactic information while training complex models towards knowledge extraction has been shown to improve performance [38]. Strubell et al. [39] have improved semantic role labelling via syntax-aware LLMs, and Jafari et al. [40] have injected syntactic features as additional embeddings into an LLM used for semantic relation extraction. One way to increase an LLM's awareness of syntactic structure is by adding local syntactic attention [see 41]. The underlying idea is to start the LLM's training from dependency parsed data and to obtain a notion of locality in the added attention based on distances in the dependency trees. In a similar direction, Bai et al. [42] modify the attention typology of the Transformer architecture based on the syntactic structure of the training data. Such approaches require a modification of the Transformer architecture. To avoid this, Zhang et al. [43] propose to syntactically enhance the LLM via specific learning objectives, more concretely syntax-guided contrastive learning. Based on syntactic structure , specific syntactic objectives are designed towards which the LLM is optimized in pre-training. A study on consistency of LLMs [17] concludes that LLMs produce inconsistent results when prompted with *syntax-preserving but differently phrased* prompts, and *different-syntax but similar semantics* prompts, suggesting that the LLMs are not suitable for extracting factual knowledge robustly and that the syntax of prompts also plays a key role. Our work on the impact of prompt syntax on rKR from transformer-based language models [29] also presents similar findings.

Biases in knowledge retrieval evaluation. LLMs may exhibit various types of biases; representation of the majority viewpoint being a common issue due to distributions prevalent within pretraining data [44], neglecting disagreements among multiple viewpoints (e.g. by majority voting) [45]. Prior works investigate individual factors (such as frequency) or LLM biases in other tasks [46], as well as knowledge retrieval [26]. With respect to the interpretation, reliability, and generalisability of knowledge retrieval, several studies [31, 47] investigate whether LLMs actually learn transferable generalisations or only exploit incidental shortcuts in the data. Cao et al. [47] explore biases in three different knowledge retrieval paradigms, namely *prompt-based retrieval*, *case-based analogy*, *context-based inference*, finding that decent performance of existing knowledge retrieval baselines tends to be driven by biased prompts that overfit to artefacts in the data, guide the LLM towards correct entity types or unintentionally leak correct answers or additional constraints applicable to the correct answer. In a similar context, Du et al. [48] discusses the shortcut learning behaviour arising due to skewed training datasets, the model, or the fine-tuning process. Schramowski et al. [49] demonstrate an intriguing similarity between human cognitive biases and those exhibited by LLMs. Using insights from psychology, they analyse the learning and decision-making processes of black-box models to reveal their biases towards right-and-wrong for decision-making. Therefore, rigorous assessment of existing benchmark datasets is necessary for generalizable insights about knowledge retrieval and inference performance, and to facilitate efficient, unbiased knowledge retrieval from LLMs.

Prompt Engineering for Knowledge Retrieval. Cao et al. [47] proposed three paradigms for factual knowledge extraction from LLMs: prompt-based, case-based, and context-based. Results suggest prompt-based retrieval is biased towards prompt structure. Prompt engineering [50] aims to create prompts that efficiently elicit desired responses from LLMs for a specific task. However, a limited number of manually created prompts only reveal a portion of the model's encoded knowledge [51], as the response can be influenced by the phrasing of the question. Thus, prompt engineering is a crucial part of knowledge retrieval from LLMs. LPAQA uses an automated mining-based and paraphrasing-based method to generate high-quality diverse prompts, as well as ensemble methods to combine answers from different prompts [51]. Automatic Prompt Engineer, proposed by [7] uses LLM models like InstructGPT [6] and instruction induction [52] to generate instruction candidates which are then improved by proposing semantically similar instruction variants to achieve human-level performance. Zhou et al. [7] investigate the ability of LLMs, such as GPT-3, to generate high-quality prompts for a variety of tasks. Initial experiments on the role of syntax in knowledge retrieval [29] find a strong interaction between prompt structure and knowledge retrieval performance.

3. Towards Syntax-aware LLM Pretraining and Prompt Engineering for Knowledge Retrieval

3.1. Research Directions & Objectives

To summarise, prior works have shown that relational knowledge is captured by LLMs to a certain extent. However, there is still insufficient understanding of how performance differs across different kinds of knowledge or relations, for instance, commonsense knowledge compared to entity-centric encyclopedic facts or transversal versus hierarchical relations. Most importantly though, the relation between, on the one hand, syntax of both pretraining corpora and prompts, and on the other hand, rKR performance, is not well understood.

Therefore, we argue that further research should be dedicated towards the following objectives and research questions.

O1. Understanding relational rKR from LLMs and the impact of syntax. Further research is required to provide a thorough understanding of relational rKR performance, inherent biases, and the impact of syntactic characteristics of both pretraining corpora and probing techniques in that context. Since prior works [31, 47] have shown that widely used rKR benchmarks may take advantage of incidental shortcuts and spurious signals in the data and thus, provide misleading insights about learned generalisations, research needs to investigate such dependencies and derive reliable probes for understanding rKR performance in LLMs. Specifically, the following research questions should be in focus:

- **[RQ1.1]** What biases can be observed in rKR from LLMs, and how are these influenced by training corpora, learning paradigms or model architectures?
- **[RQ1.2]** Which factors impact the reliability and meaningfulness of experimental rKR results?
- **[RQ1.3]** What impact do explicit and implicit syntactic features of prompts and pretraining corpora have on the LLM-based inference of relational knowledge?

O2. Improving inference of relational knowledge through syntax-aware LLM pretraining and probing and through injecting additional symbolic knowledge. Given the capabilities of self-supervised LLMs for rKR and inference, there is significant potential to exploit LLMs as part of various NLP tasks that traditionally had to resort to costly supervised approaches, such as knowledge base construction or question answering. Hence, building on the insights from O1 it is feasible to derive strategies that exploit syntactic features for improving the rKR and inference capacities of LLMs. These may comprise syntax-informed pretraining strategies, LLM training paradigms, and prompt engineering. Specifically, we consider the following research questions:

- **[RQ2.1]** How can pretraining of neural LLMs be informed by syntax to improve the inference of relational knowledge?
- **[RQ2.2]** How can syntax-informed prompting or verbalisation strategies improve relational knowledge extraction from LMs?
- **[RQ2.3]** How can we exploit symbolic knowledge to improve relational knowledge extraction from syntax-aware LLMs?

Research geared towards addressing these questions will improve reusable approaches that exploit syntax to optimise LLMs and probing techniques towards the rKR task. By advancing the understanding of the interplay of semantics and syntax in pretrained LLMs, such research also facilitates computationally less expensive training paradigms that preserve the rKR capacities of larger models while requiring fewer parameters.

3.2. Preliminary Analysis

Paraphrasing a prompt may introduce a variety of changes, including semantic ones that change the information content of the prompt as well as syntactic ones that merely change the form in which the same content is expressed. Our previous work [29] studied the impact of prompt syntax on the rKR capacity of LLMs. We expanded the well-known and commonly used T-REx subset of the LAMA-probe [16]. We used a template-based approach to paraphrase simple LAMA prompts into more complex grammatical structure. We then analyse the LLM performance for these structurally different but semantically equivalent prompts. Our preliminary study revealed that simple prompts work better than complex forms of sentences. Furthermore, we observed that the performance across the syntactical variations for simple relations better as compared to complex relations. Our study showed that LLMs indeed struggle to generalise knowledge across grammatical structures, highlighting the relationship between syntax and semantics in the context of rKR through LLMs.

4. Conclusions & Outlook

In this position paper, we have laid out the motivation and future directions for research concerned with investigating the impact of syntax on knowledge retrieval from LLMs. Building on observations that LLMs learn syntax to a certain extent and that prompt syntax impacts knowledge retrieval performance [29], we argue that understanding the impact of syntax on knowledge retrieval performance is a crucial prerequisite for understanding how LLMs learn language representations. In addition, a deeper understanding of the impact of syntax of prompts and pretraining corpora will facilitate more efficient knowledge retrieval from LLMs. Given the deficiencies of current rKR benchmarks, research in this area has to invest in creating more controlled benchmark probes able to isolate the effects of syntax on rKR performance, as well as pretraining corpora and paradigms where the amount of information and prevalent syntax can be controlled rigorously. Moreover, research into injecting syntactic knowledge from supervised dependency parsing into LLMs is also a promising avenue for improving the LLMs' rKR performance.

References

[1] B. Fetahu, U. Gadiraju, S. Dietze, Improving entity retrieval on structured data, in: The Semantic Web - ISWC 2015, Springer International Publishing, Cham, 2015, pp. 474–491.

[2] R. Yu, U. Gadiraju, B. Fetahu, et al., Knowmore - knowledge base augmentation with structured web markup., Semantic Web 10 (2019) 159–180.

[3] N. Tempelmeier, E. Demidova, S. Dietze, Inferring missing categorical information in noisy and sparse web markup, in: Proceedings of The Web Conference 2018 (WWW 2018), 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, et al., Bert: Bidirectional encoder representations from transformers (2016).

[5] A. Radford, J. Wu, R. Child, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[6] L. Ouyang, J. Wu, X. Jiang, et al., Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155 (2022).

[7] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language

models are human-level prompt engineers, 2023. URL: https://arxiv.org/abs/2211.01910. arXiv:2211.01910.

[8] B. Heinzerling, K. Inui, Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1772–1791. URL: https://aclanthology.org/2021.eacl-main.153. doi:10.18653/v1/2021.eacl-main.153.

[9] D. Sachan, Y. Zhang, P. Qi, et al., Do syntax trees help pre-trained transformers extract information?, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2647–2661. doi:10.18653/v1/2021.eacl-main.228.

[10] A. Tamborrino, N. Pellicano, B. Pannier, et al., Pre-training is (almost) all you need: An application to commonsense reasoning, arXiv preprint arXiv:2004.14074 (2020).

[11] J. Hu, J. Gauthier, P. Qian, et al., A systematic assessment of syntactic generalization in neural language models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1725–1744. doi:10.18653/v1/2020.acl-main.158.

[12] J. Hewitt, C. D. Manning, A structural probe for finding syntax in word representations, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4129–4138. URL: https://aclanthology.org/N19-1419. doi:10.18653/v1/N19-1419.

[13] D. Arps, Y. Samih, L. Kallmeyer, H. Sajjad, Probing for constituency structure in neural language models, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6738–6757. URL: https://aclanthology.org/2022.findings-emnlp.502.

[14] G. S. Halford, W. H. Wilson, S. Phillips, Relational knowledge: The foundation of higher cognition, Trends in cognitive sciences 14 (2010) 497–505.

[15] L. Ouyang, J. Wu, X. Jiang, et al., Aligning language models to follow instructions, ???? URL: https://openai.com/research/instruction-following.

[16] F. Petroni, T. Rocktäschel, S. Riedel, et al., Language models as knowledge bases?, 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), ACL, 2019.

[17] Y. Elazar, N. Kassner, S. Ravfogel, et al., Measuring and improving consistency in pretrained language models, Transactions of the Association for Computational Linguistics 9 (2021) 1012–1031.

[18] B. Heinzerling, K. Inui, Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries, arXiv preprint arXiv:2008.09036 (2020).

[19] B. Cao, H. Lin, X. Han, et al., Knowledgeable or educated guess? revisiting language models as knowledge bases, arXiv preprint arXiv:2106.09231 (2021).

[20] F. Petroni, P. Lewis, A. Piktus, et al., How context affects language models' factual predictions, arXiv preprint arXiv:2005.04611 (2020).

[21] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, H. Chen, Knowprompt:

Knowledge-aware prompt-tuning with synergistic optimization for relation extraction, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2778–2788.

[22] L. Pandia, A. Ettinger, Sorting through the noise: Testing robustness of information processing in pre-trained language models, arXiv preprint arXiv:2109.12393 (2021).

[23] N. Kassner, H. Schütze, Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020. URL: https://www.aclweb.org/anthology/2020.acl-main.698.

[24] A. Ravichander, E. Hovy, K. Suleman, A. Trischler, J. C. K. Cheung, On the systematicity of probing contextualized word representations: The case of hypernymy in bert, in: Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, 2020, pp. 88–102.

[25] J. D. Hwang, C. Bhagavatula, R. Le Bras, et al., (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 6384–6392.

[26] Z. Zhong, D. Friedman, D. Chen, Factual probing is [MASK]: Learning vs. learning to recall, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

[27] J.-C. Kalo, L. Fichtel, Kamel: Knowledge analysis with multitoken entities in language models, in: Proceedings of the Conference on Automated Knowledge Base Construction, 2022.

[28] D. Gao, Y. Jia, L. Li, et al., Kmir: A benchmark for evaluating knowledge memorization, identification and reasoning abilities of language models, arXiv preprint arXiv:2202.13529 (2022).

[29] S. Linzbach, T. Tressel, L. Kallmeyer, S. Dietze, H. Jabeen, Decoding prompt syntax: Analysing its impact on knowledge retrieval in large language models, in: Natural Language Processing for Knowledge Graph Creation NLP4KGC, Workshop at The Web Conference WWW'23, 2023.

[30] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, arXiv preprint arXiv:2212.10403 (2022).

[31] R. Branco, A. Branco, J. António Rodrigues, et al., Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 1504–1521.

[32] Y. Goldberg, Assessing bert's syntactic abilities, arXiv preprint arXiv:1901.05287 (2019).

[33] Y. Lin, Y. C. Tan, R. Frank, Open sesame: Getting inside BERT's linguistic knowledge, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Florence, Italy, 2019, pp. 241–253. URL: https://www.aclweb.org/anthology/W19-4825. doi:10.18653/v1/W19-4825.

[34] X. Zhu, L. Zhu, J. Guo, et al., Gl-gcn: Global and local dependency guided graph convolutional networks for aspect-based sentiment classification, Expert Syst. Appl. 186 (2022). doi:10.1016/j.eswa.2021.115712.

[35] X. Zhu, G. Tang, P. Wang, et al., Dynamic global structure enhanced multi-channel

graph neural network for session-based recommendation, Information Sciences 624 (2023) 324–343.

[36] L. Kallmeyer, B. QasemiZadeh, J. C. Cheung, Coarse lexical frame acquisition at the syntax–semantics interface using a latent-variable PCFG model, in: Proceedings of *SEM 2018, 2018, pp. 130–141.

[37] B. QasemiZadeh, M. R. L. Petruck, R. Stodden, et al., SemEval-2019 task 2: Unsupervised lexical frame induction, in: Proceedings of the 13th International Workshop on Semantic Evaluation, ACL, Minneapolis, Minnesota, USA, 2019, pp. 16–30.

[38] D. Sundararaman, V. Subramanian, G. Wang, et al., Syntactic knowledge-infused transformer and bert models, in: CEUR Workshop Proceedings, volume 3052, CEUR Workshop Proceedings, 2021.

[39] E. Strubell, P. Verga, D. Andor, et al., Linguistically-informed self-attention for semantic role labeling, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 5027–5038.

[40] M. M. Jafari, S. Behmanesh, A. Talebpour, et al., Improving pre-trained language model for relation extraction using syntactic information in persian, in: Proceedings of The Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021, Association for Computational Linguistics, Trento, Italy, 2021, pp. 38–44.

[41] Z. Li, Q. Zhou, C. Li, et al., Improving BERT with syntax-aware local attention, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 645–653.

[42] J. Bai, Y. Wang, Y. Chen, et al., Syntax-BERT: Improving pre-trained transformers with syntax trees, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3011–3020.

[43] S. Zhang, W. Lijie, X. Xiao, et al., Syntax-guided contrastive learning for pre-trained language model, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2430–2440.

[44] E. M. Bender, T. Gebru, A. McMillan-Major, et al., On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[45] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110.

[46] R. Mao, Q. Liu, K. He, et al., The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, IEEE Transactions on Affective Computing (2022) 1–11. doi:10.1109/TAFFC.2022.3204972.

[47] B. Cao, H. Lin, X. Han, L. Sun, L. Yan, M. Liao, T. Xue, J. Xu, Knowledgeable or educated guess? revisiting language models as knowledge bases, 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), Association for Computational Linguistics, Online, 2021.

[48] M. Du, F. He, N. Zou, et al., Shortcut learning of large language models in natural language understanding: A survey, arXiv preprint arXiv:2208.11857 (2022).

[49] P. Schramowski, C. Turan, N. Andersen, et al., Large pre-trained language models contain human-like biases of what is right and wrong to do, Nature Machine Intelligence 4 (2022) 258–268.

[50] S. H. Bach, V. Sanh, Z.-X. Yong, A. Webson, C. Raffel, N. V. Nayak, A. Sharma, T. Kim, M. S. Bari, T. Fevry, et al., Promptsource: An integrated development environment and repository for natural language prompts, 2022.

[51] Z. Jiang, F. F. Xu, J. Araki, et al., How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438. doi:10.1162/tacl_a_00324.

[52] O. Honovich, U. Shaham, S. R. Bowman, et al., Instruction induction: From few examples to natural language task descriptions, arXiv preprint arXiv:2205.10782 (2022).