预训练语言模型中的知识分析、萃取与增强

陈玉博[1,2], **曹鹏飞**[1,2], **王晨皓**[1,2], **李嘉淳**[1,2], **刘康**[1,2,3], **赵军**[1,2]
[1]中国科学院自动化研究所复杂系统认知与决策实验室
[2]中国科学院大学人工智能学院
[3]北京智源人工智能研究院
{yubo.chen, pengfei.cao, chenhao.wang, kliu, jzhao}@nlpr.ia.ac.cn

摘要

近年来，大规模预训练语言模型在知识密集型的自然语言处理任务上取得了令人瞩目的进步。这似乎表明，预训练语言模型能够自发地从语料中学习大量知识，并隐式地保存在参数之中。然而，这一现象的背后机理仍然萦绕着许多谜团，语言模型究竟掌握了哪些知识，如何提取和利用这些知识，如何用外部知识弥补模型不足，这些问题都亟待进一步探索。在本次讲习班中，我们将重点介绍在预训练语言模型知识分析、知识萃取、知识增强等领域的近期研究进展。

关键词： 预训练语言模型 ； 知识分析 ； 知识萃取 ； 知识增强

时长： 90分钟

目标听众： 目标听众主要为自然语言处理领域和知识图谱领域的研究人员和工程人员，听众将在本次讲习班中了解预训练语言模型相关的知识分析、知识萃取、知识增强等领域的最新研究进展。

内容大纲：

- 预训练语言模型简介（5分钟）
- 预训练语言模型的知识分析（包括知识的探测、定位和编辑）（40分钟）
- 预训练语言模型的知识萃取（从预训练语言模型提取符号知识）（15分钟）
- 预训练语言模型的知识增强（用外部知识辅助预训练语言模型）（30分钟）

Knowledge Analysis, Extraction and Enhancement in Pre-trained Language Models

Yubo Chen[1,2], **Pengfei Cao**[1,2], **Chenhao Wang**[1,2], **Jiachun Li**[1,2], **Kang Liu**[1,2,3], **Jun Zhao**[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Beijing Academy of Artificial Intelligence
{yubo.chen, pengfei.cao, chenhao.wang, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Recently, large-scale pre-trained language models have made remarkable progress in knowledge-intensive natural language processing tasks. It seems to indicate that pre-trained language models can naturally learn extensive knowledge from the corpus and implicitly encode it in the parameters. However, the underlying mechanisms behind the phenomenon remain largely unknown. Questions such as what knowledge has been acquired by language models, how to extract and utilize the knowledge, and how external knowledge can be incorporated to address the limitations of models, are

第二十二届中国计算语言学大会论文集，第1页–第8页，哈尔滨，中国，2023年8月3日至5日。

all awaiting further exploration. In this tutorial, we will focus on introducing recent research advancements in the knowledge analysis, knowledge extraction, and knowledge enhancement of pre-trained language models.

Keywords: Pre-trained Language Models , Knowledge Analysis , Knowledge Extraction , Knowledge Enhancement

Duration: 90 minutes

Targeted Audience: The target audiences are researchers and engineers in the field of natural language processing and knowledge graph. In this tutorial, the audiences will learn about the latest research progress in knowledge analysis, knowledge extraction, and knowledge enhancement related to pre-trained language models.

Outline:

- Introduction of Pre-trained Language Models (5 minutes)
- Knowledge Analysis in PLMs (Knowledge Probing, Locating and Editing)（40 minutes）
- Knowledge Extraction in PLMs (Extracting Symbolic Knowledge from PLMs)（15 minutes）
- Knowledge Enhancement in PLMs (Assisting PLMs with External Knowledge) (30 minutes)

1 内容介绍

近年来，预训练语言模型已逐渐成为自然语言处理领域的基座模型。相关实验现象表明，预训练语言模型能够自发地从预训练语料中学到一定的语言学知识、世界知识和常识知识，从而在知识密集型任务上获得出色的表现(AlKhamissi et al., 2022; Safavi and Koutra, 2021; Petroni et al., 2019)。然而，预训练语言模型中的知识隐式地存储在参数之中，难以显式地对预训练语言模型中的知识进行分析和利用。同时，预训练语言模型在知识和推理上的表现并不可靠，常常会出现“幻觉”现象(Ji et al., 2022)，给出与知识冲突的预测结果。这些因素阻碍了预训练语言模型提供可靠的知识服务。因此，探究模型掌握知识的机理、研究如何提取和补充语言模型中的知识成为近期的研究热点。

本次讲习班主要内容包括预训练语言模型中的知识分析、预训练语言模型的知识萃取、知识增强的预训练语言模型三个部分，听众将在本次讲习班中了解到近期研究中对预训练语言模型掌握知识情况的认识、从预训练语言模型中提取符号知识的实现方案、利用外部知识增强模型弥补缺陷的各类方法。

1.1 知识分析

预训练语言模型在预训练阶段学习了大量语料，并隐式地从中获取了多种类型的知识。然而，这些知识的存在状态对人类来说是不透明的。我们很难直接确定预训练语言模型掌握了哪些知识，也并不清楚确定这些知识所在的位置和访存机制。因此，想要更好地研究预训练语言模型中的知识，势必要对语言模型进行深入的知识分析。

首先，为了确定预训练语言模型中存在哪些知识，需要进行最基本的知识探测工作。知识探测的基本思想是用已经整理好的人类知识，检验模型掌握的程度。目前，互联网上公开可用的知识资源可以大致分为语言学知识、世界知识、常识知识三类(Wang et al., 2021a)。对于这三类知识，现在已经有一些在预训练语言模型之上的探测分析工作，可以大体分为有训练方法和无训练方法。语言学知识方面，相关工作主要使用语言模型的隐层表示进行语言结构预测，检查对语言结构知识的掌握程度(Liu et al., 2019)；或直接通过提示模型进行生成，探测模型是否掌握词汇关系知识(Jain and Anke, 2022)。世界知识方面，相关工作主要依托现有三元组形式的知识库构造知识补全任务，并通过特定提示词引导模型预测，探测语言模型正确预测结果的能力(Petroni et al., 2019; Jiang et al., 2020; Shin et al., 2020; Liu et al., 2021a)。常识知识方

面，相关工作主要依靠打分对比判断的形式，分析模型能否正确区分句子是否符合常识(Zhou et al., 2020; Li et al., 2022)。这些研究工作表明，语言模型一定程度上拥有不同种类的知识。但是，由于语言模型知识探测结果可能受到多种因素干扰，目前的研究尚不足以可靠地确定语言模型拥有知识的范围，仍然需要更有效的实验设计(Cao et al., 2021a)。

除了从整体上探测语言模型掌握知识的能力表现。另一类语言分析研究试图从语言模型结构中定位出与特定类型知识相关的部分，从而更深入地理解语言模型访存知识的机制。目前在这方面最主要的假设是语言模型各层的前馈网络模块起到了类似键值存储的作用(Geva et al., 2021)。有关研究证明了不同层的前馈网络模块能识别不同程度的语义信息，并且前馈网络的隐层表示、变换矩阵参数与世界知识能否正确补全有较强的因果联系(Dai et al., 2022; Meng et al., 2022a)。这些研究初步实现了对部分知识的存储位置定位，从整体角度说明了在语言模型中，特定类型的知识可能存储于局部参数之中，调整这些参数能对语言模型的知识表现产生影响。

此外，预训练语言模型的知识来自于训练语料，并固化在模型参数之中。尽管模型可以通过训练获取新的知识，但这一过程往往会引入对旧知识的灾难性遗忘，难以控制。因此，在分析预训练语言模型的知识范围、存储位置基础上，另一类最新研究正在探索定向编辑语言模型中的知识。这些工作旨在有效编辑目标知识的同时，尽量保持无关知识不受到影响，大体上可以分为超网络方法和定向知识编辑方法。前者主要依靠数据驱动方法和元学习思想，训练一个根据知识编辑内容产生模型更新参数的超网络，从而满足知识编辑的优化目标(Cao et al., 2021b; Mitchell et al., 2022)。后者主要在知识定位分析的基础上，确定知识存储形式和预期更新结果，然后通过局部更新的方法实现知识的有效编辑(Dai et al., 2022; Meng et al., 2022a)。目前，知识编辑的研究正在逐渐扩大规模(Meng et al., 2022b)，但是仍然存在知识类型有限、难以连续更新等问题。

1.2 知识萃取

预训练语言模型中蕴含着大量知识，但这些知识隐式地存储在模型参数之中，难以直接访问和量化分析。此外，目前构造结构化知识库是一项耗时耗力的任务，语料中蕴含的许多知识可能至今尚未得到结构化组织。因此，随着预训练语言模型的知识能力不断进步，越来越多的研究工作试图将语言模型中的隐式知识萃取出来，得到符号化显式表达的知识，用于进一步分析和应用。

由于常识知识收集难度高，且现有知识资源严重不足，因此当下的知识萃取工作更多在常识知识上开展。这些方法主要将知识萃取过程分解成多个子任务，每个子任务依靠提示引导预训练语言模型进行生成或判别，从而获取大量结构化知识候选。这些结构化知识候选再通过进一步的过滤得到高质量的知识集合。在模型基础能力强，萃取流程设计合理的情况下，预训练语言模型能够用于产生质量与人类相当的常识知识(West et al., 2022)。最新的知识萃取方法正在逐渐放宽对预训练语言模型要求的条件(Wang et al., 2022; Bhagavatula et al., 2022)，并且将萃取实践扩展到世界知识在内的更多知识类型上(Cohen et al., 2023)。

1.3 知识增强

近些年来，预训练语言模型暴露出来的推理能力不足、产生幻觉等现象愈发受到人们重视(Ji et al., 2022)。知识增强作为缓解这些问题的重要方法，在近年来有一系列深入研究。

第一类知识增强方法主要从预训练目标任务设计的角度考虑问题。由于早期预训练语言模型大多基于掩码重建任务进行预训练，所以采用知识引导的掩码策略来引入知识成为一种较为直接的知识增强方式。例如，百度ERNIE(Sun et al., 2019)引入了短语级和实体级的掩码粒度，将额外的知识融入到掩码语言模型任务学习中。Graph-Guided MLM(Shen et al., 2020)采用了一种知识图谱引导的掩码策略，筛选出信息量更高的实体以高效地学习KG中的结构化知识。除此之外，一些工作还提出了更多样化的预训练任务，如KALM工作中的Knowledge-Aware任务(Rosset et al., 2020)、KEPLER中的知识表示损失(Wang et al., 2021c)、百度ERNIE 3.0工作中的多层次知识相关任务等(Sun et al., 2021)。

第二类知识增强方法主要从模型结构设计的角度考虑问题。在现有基础模型之上增加知识注入的相关模块，以显式的方式将知识融合入模型中。例如通过修改模型输入端设计，增加新的Knowledge Embedding层，在输入端将知识结构表示与原有的文本信息结合编码，典型工作包括K-BERT(Liu et al., 2020)、CoLAKE(Sun et al., 2020)等。另一些工作则是引入额外

的模块编码知识信息，并通过信息交互融合模块在深层将文本和知识信息融合，典型工作包括清华ERNIE(Zhang et al., 2019)，KG-BART(Liu et al., 2021b)，KnowBERT(Peters et al., 2019)等。

第三类知识增强方法主要从外部模块交互的角度考虑问题。上述两类方法都改变了预训练语言模型原有的任务或结构，需要针对性的训练，从而将外部知识注入到模型之中。随着模型的复杂化和参数量的增加，其灵活性欠缺和训练成本偏高的缺点也逐渐被放大，这催生了基于外部模块交互的知识增强方法。此类方法在尽量不修改基础模型的前提下，开发与基础模型解耦的知识增强模块，通过优化该模块实现知识增强，具有成本低、可扩展等特点。例如，K-ADAPTER(Wang et al., 2021b)将小型的知识模型以插件的形式连接到语言模型上，通过小模型的学习向其中注入特定知识。KGLM(IV et al., 2019)借助动态的本地小型知识图谱，在生成阶段不断补充相关知识。此外，最新研究也逐渐扩展到GPT-3、ChatGPT这类千亿参数的超大规模语言模型上，通过微调或提示设计等方式，让语言模型学会生成操作指令，与搜索引擎、数据库等外部知识工具交互，从中得到与输入文本相关的知识，作为模型的补充输入解决知识相关的任务(Nakano et al., 2021; Peng et al., 2023)。

2 推荐阅读列表

本次讲习班主要涉及预训练语言模型背景下的知识工程研究，在相关领域有一些优秀的综述性文章。

1. 预训练语言模型用于知识密集型自然与语言处理任务(Yin et al., 2022)
2. 预训练语言模型的工作机理、知识分析(Rogers et al., 2020; AlKhamissi et al., 2022)
3. 预训练语言模型中的知识探测、增强方法(Safavi and Koutra, 2021)
4. 融合外部知识资源的技术综述(Wang et al., 2023)

3 讲者介绍

陈玉博，中国科学院自动化研究所副研究员，研究方向为自然语言处理和知识图谱，在ACL、EMNLP、AAAI 等国际重要会议和期刊发表学术论文40 余篇，其中多篇论文入选Paper Digest最具影响力论文，曾获多次最佳论文奖（NLP-NABD 2016、CCKS 2017、CCL 2020、CCKS 2020），Google Scholar引用量4100余次。出版学术专著两部《知识图谱》、《知识图谱：算法与实践》，由人工智能学会推荐入选十三五国家重点图书出版规划教材，连续多年在中国科学院大学主讲《知识图谱》课程，2021 年获得中国科学院大学优秀课程。主持国家自然科学基金面上项目、青年基金项目，参与国家自然科学基金重点项目、2030新一代人工智能重大项目、重点研发计划课题。主持研发的信息抽取和知识图谱构建系统多次获得国际/国内学术评测冠亚军。入选2020 年第五届中国科协青年人才托举工程、2022 年全球华人AI 青年学者、2022 年中国科学院青年创新促进会会员、2022北京智源人工智能青年科学家俱乐部，担任中国中文信息学会青年工作委员会秘书长、COLING 2022领域主席、Data Intelligence编委等。获2018 年中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖（个人排名第四），2019 年度北京市科学技术进步奖一等奖（个人排名第五）。

4 其他贡献者介绍

本讲习班的其它贡献者有：

赵 军，中国科学院自动化研究所研究员，博士生导师；中国科学院大学人工智能学院岗位教授。研究领域为自然语言处理、知识图谱、信息抽取、问答系统、大模型等。作为项目负责人承担国家自然科学基金重点项目、科技创新2030-新一代人工智能重大项目等多项国家级重要科研项目以及企业应用项目。在ACL、IJCAI、SIGIR、AAAI、COLING、EMNLP、TKDE等顶级国际会议和重要学术期刊上发表论文100余篇，曾获第25届国际计算语言学大会COLING 2014最佳论文奖，Google Scholar引用量19000余次。出版学术专著两部《知识图谱》、《知识图谱：算法与实践》，由人工智能学会推荐入选十三五国家重点图书出版规划教材，连续多年在中国科学院大学主讲

《知识图谱》课程，2021 年获得中国科学院大学优秀课程，获朱李月华优秀教师奖。主持研发的“大规模开放域文本知识获取与应用平台”获得2018年中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖，主持完成的“大规模知识图谱构建关键技术与应用”项目获得2019年度北京市科学技术进步奖一等奖。兼任中国中文信息学会常务理事，语言与知识计算专委会副主任，《中文信息学报》编委，Machine Intelligence Research （MIR）编委等学术职务。

刘康，中国科学院自动化研究所研究员、博士生导师，中国科学院大学岗位教授，北京智源人工智能研究院青年科学家。研究领域包括自然语言处理、文本信息抽取、知识图谱、问答系统等。在自然语言处理、知识工程等领域国际重要会议和期刊发表多篇学术论文，Google Scholar引用16000余次，单篇引用数达到2700余次，H-Index为50。2020-2023连续入选Aminer ”AI 2000人工智能全球最具影响力提名学者。曾获COLING 2014最佳论文奖、Google Focused Research Award（2015、2016）、中国中文信息学会“汉王青年创新一等奖”、中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖、北京市科学技术进步一等奖等多项学术奖励。2019年获得国家自然科学基金委优秀青年基金支持，2020年入选中国科学院青年创新促进会优秀会员。目前兼任中国中文信息学会理事、中国中文信息学会计算语言学专委会、中国中文信息学会语言与知识计算专委会秘书长等学术职务。目前担任Pattern Recognition、TACL等学术期刊编委，也曾任ACL、AAAI、EMNLP、CIKM、ISWC、EACL等国际高水平学术会议（Senior）Area Chair/Senior PC member。

曹鹏飞，中国科学院自动化研究所助理研究员。主要研究方向为自然语言处理、知识图谱、信息抽取。以一作身份在AAAI、ACL、EMNLP等人工智能领域国际顶级学术会议上发表多篇论文。曾任中国中文信息学会青年工作委员会学生执委会的执行委员，中国自然语言处理学生研讨会的博士生论坛主席，并担任TKDE、ACL、EMNLP、AAAI等著名国际期刊和学术会议的审稿人。

王晨皓，中国科学院自动化研究所2019级硕博生。主要研究方向为知识图谱、常识知识获取与知识探测、常识推理。并在AAAI、EMNLP、CCKS等人工智能领域学术会议发表多篇论文。

李嘉淳，中国科学院自动化研究所2022级直博生。主要研究方向为常识知识获取与知识探测。

参考文献

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *CoRR*, abs/2204.06031.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2022. I2D2: inductive knowledge distillation with neurologic and self-imitation. *CoRR*, abs/2212.09246.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021a. Knowledgeable or educated guess? revisiting language models as knowledge bases. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1860–1874. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021b. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. *CoRR*, abs/2301.12810.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Robert L. Logan IV, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5962–5971. Association for Computational Linguistics.

Devansh Jain and Luis Espinosa Anke. 2022. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction. In Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, José Camacho-Collados, and Alessandro Raganato, editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2022, Seattle, WA, USA, July 14-15, 2022*, pages 151–156. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11838–11855. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021a. GPT understands, too. *CoRR*, abs/2103.10385.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021b. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6418–6425. AAAI Press.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual knowledge in GPT. *CoRR*, abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *CoRR*, abs/2210.07229.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Trans. Assoc. Comput. Linguistics*, 8:842–866.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Knowledge-aware language model pretraining. *CoRR*, abs/2007.00655.

Tara Safavi and Danai Koutra. 2021. Relational world knowledge representation in contextual language models: A review. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1053–1067. Association for Computational Linguistics.

Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8980–8994. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4222–4235. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3660–3670. International Committee on Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137.

Chenhao Wang, Yubo Chen, Zhipeng Xue, Yang Zhou, and Jun Zhao. 2021a. Cognet: Bridging linguistic knowledge, world knowledge and commonsense knowledge. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial*

Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 16114–16116. AAAI Press.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021b. K-adapter: Infusing knowledge into pre-trained models with adapters. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1405–1418. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021c. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguistics*, 9:176–194.

Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2022. Cn-automic: Distilling chinese commonsense knowledge from pretrained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9253–9265. Association for Computational Linguistics.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, Suparna De, and Amir Hussain. 2023. Fusing external knowledge resources for natural language understanding techniques: A survey. *Inf. Fusion*, 92:190–204.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Da Yin, Li Dong, Hao Cheng, Xiaodong Liu, Kai-Wei Chang, Furu Wei, and Jianfeng Gao. 2022. A survey of knowledge-intensive NLP with pre-trained language models. *CoRR*, abs/2202.08772.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press.