A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models

Da Yin[1*], **Li Dong**[2], **Hao Cheng**[2], **Xiaodong Liu**[2], **Kai-Wei Chang**[1], **Furu Wei**[2], **Jianfeng Gao**[2]
[1]University of California, Los Angeles
[2]Microsoft Research
{da.yin, kwchang}@cs.ucla.edu, {lidong1, chehao, xiaodl, fuwei, jfgao}@microsoft.com

Abstract

Despite the dramatic increase of model capability, large-scale pre-trained language models (PLMs) are still insufficient to handle a wide range of knowledge-intensive natural language processing (NLP) tasks, such as open-domain question answering and multi-turn goal-directed dialog, mainly because these models, trained only on raw text corpora, cannot effectively leverage external knowledge, e.g., commonsense and encyclopedic knowledge, for problem solving. To address the problem, there is a growing interest in augmenting PLMs by incorporating various types of external knowledge in model training and inference. In this paper, we review the recent advances in pre-trained language model-based knowledge-enhanced models (PLMKEs) by presenting progress that has been made and challenges still being faced on three vital elements: knowledge sources, knowledge-intensive NLP tasks, and knowledge fusion methods. We hope that the paper can provide NLP researchers with insights and directions for further research.

1 Introduction

Pre-trained language models (PLMs) [e.g., Devlin *et al.*, 2019; Radford *et al.*, 2019; He *et al.*, 2020] have achieved enormous successes in natural language processing (NLP). Trained using self-supervised language modeling objectives over massive text corpora, the resulting PLMs provide powerful text representations for supervised fine-tuning on downstream tasks, leading to new state of the arts on a wide range of NLP tasks [Wang *et al.*, 2018, 2019]. In particular, recent studies show that certain knowledge (linguistic or factual knowledge [Manning *et al.*, 2020; Petroni *et al.*, 2019; Roberts *et al.*, 2020; Dai *et al.*, 2021]) is implicitly stored in their parameters which partially explains the superior generalization abilities of PLM-based NLP models. However, the AI systems that are merely based on the *implicit knowledge* from PLMs lack the ability to leverage *explicit* encyclopedic and commonsense knowledge for problem solving, and thus are insufficient to deal with knowledge-intensive NLP tasks, such as daily news retrieval, question answering, multi-turn task-oriented dialog [e.g., Guu *et al.*, 2020; Gao *et al.*, 2022].

The growing real-world needs motivate the development of **P**re-Trained **L**anguage **M**odel-based **K**nowledge-**E**nhanced Models (**PLMKE**s). In this paper, we review the recent advances in PLMKEs for NLP. In a PLMKE, input-relevant external knowledge is fetched using a *knowledge interface* and then is used to generate the output via a *knowledge fusion module*. PLMKEs have been developed for various knowledge-intensive tasks including open-domain question answering [Guu *et al.*, 2020; Lewis *et al.*, 2020; Cheng *et al.*, 2021], fact verification [Zhou *et al.*, 2019b; Liu *et al.*, 2020], entity linking [Jiang *et al.*, 2021] and commonsense reasoning [Lin *et al.*, 2019; Yasunaga *et al.*, 2021].

Since different tasks require different types of knowledge, most recent PLMKEs customize the source for knowledge interface and the design of corresponding knowledge fusion method. Thus, we structure our survey so that different knowledge-intensive scenarios can be understood through our lens. Specifically, we focus our discussion on three vital items related to PLMKEs:

Knowledge Sources: Knowledge sources (Wikipedia, [e.g., Bollacker *et al.*, 2008; Vrandečić, 2012; Speer *et al.*, 2017; Sap *et al.*, 2019a; Zhang *et al.*, 2020a]) provide external knowledge to PLMKEs and lay the foundations of PLMKEs along with pre-trained language models. The contents of knowledge sources are also decisive to the tasks PLMKEs can solve.

Knowledge-Intensive NLP Tasks: Knowledge-intensive NLP tasks [e.g., NIST, 2004; Thorne *et al.*, 2018; Kwiatkowski *et al.*, 2019; Petroni *et al.*, 2021] are testbeds to evaluate the performance of PLMKEs, verifying whether the selected knowledge sources are appropriate and the effectiveness of knowledge fusion methods.

Knowledge Fusion Methods: Knowledge fusion methods [e.g., Zhang *et al.*, 2019; Lin *et al.*, 2019; Guu *et al.*, 2020; Sun *et al.*, 2021] exploit the use of external knowledge to enhance pre-trained language models to achieve better performance on knowledge-intensive NLP tasks.

The rest of the paper is structured as follows.

- Section 2 reviews knowledge sources used. What are the types of commonly used knowledge sources? What is the format of knowledge stored in the knowledge

[*]The work was mainly done during internship at Microsoft.

sources? What are the characteristics of the knowledge sources?

- Section 3 reviews knowledge-Intensive NLP Tasks. What are the common knowledge-intensive NLP tasks that PLMKEs are applied on? What knowledge is useful to solve the tasks?
- Section 4 reviews knowledge fusion Methods. How do we categorize the numerous knowledge fusion methods? What roles do the different categories of fusion methods typically play during knowledge fusing? What are fusion methods adopted in commonly used PLMKEs and why?
- Section 5 presents challenges and future directions.

We notice that there are two contemporaneous surveys [Wei *et al.*, 2021; Yang *et al.*, 2021] about knowledge-enhanced NLP models. However, they both focus on the models that incorporate knowledge during the pre-training stage, which is subsumed as a sub-topic of this survey. In addition, we propose a taxonomy to not only clearly categorize PLMKEs but also present new perspectives for future directions.

2 Knowledge Sources

Knowledge sources provide pre-trained language models with needed knowledge and empower them with higher capability to handle knowledge-intensive NLP tasks. We list the common knowledge sources leveraged in PLMKEs in Table 1, and further categorize them into two groups: encyclopedic knowledge and commonsense knowledge.

2.1 Encyclopedic Knowledge

Encyclopedic knowledge contains attributes (e.g., age, duration) about entities (e.g., person, event) and the relations (e.g., educated at, followed by) between entities. Wikipedia is one of the prevalent knowledge sources providing massive amount of encyclopedic knowledge including biography of a person and background of an event. Extracted from unstructured text corpora, factual knowledge bases aim at uncovering graph structures of entities and converting them into structured database. Typically, structured encyclopedic knowledge is represented by triplets containing entity names and their relationships (e.g., <Tom Hanks, occupation, actor>). Factual knowledge bases (e.g., Wikidata) are also widely used in PLMKEs [Peters *et al.*, 2019; Agarwal *et al.*, 2021].

2.2 Commonsense Knowledge

Commonsense knowledge includes the basic facts about situations in human's daily life. It involves everyday events and their effects (e.g., mop up the floor if we split food over it), facts about beliefs and desires (e.g., study hard to win scholarship), and properties of objects (e.g., goat has four legs). Thus, different from encyclopedic knowledge, commonsense knowledge is usually shared by most people and implicitly assumed in communications.

Similar to the storage method of factual knowledge bases, commonsense knowledge sources also use triplets to represent knowledge. These sources depict commonsense including subtype relationship between objects (e.g., <apple, IsA, fruit>in ConceptNet), cause and effect of an event (e.g., <personX adopts a pet, Effects, play with the pet>), and intent of human behaviour (e.g., <personX adopts a pet, Causes, to have a companion>). We can observe that the main difference from factual knowledge bases is that the triplets contain everyday objects and their elements are typically described with a short sentence. Recent PLMKEs [Lin *et al.*, 2019; Mitra *et al.*, 2019] mostly utilize the sources including ConceptNet and ATOMIC as external knowledge to enhance models' commonsense reasoning capacity.

Knowledge Types	Knowledge Sources	Knowledge Domains
Encyclopedic Knowledge	Wikipedia/Wikidata Vrandečić [2012]	open domain
	DBPedia Auer *et al.* [2007]	
	Freebase Bollacker *et al.* [2008]	
	UMLS Bodenreider [2004]	biomedicine
	AMiner Tang *et al.* [2008]	science
Commonsense Knowledge	ConceptNet Speer *et al.* [2017]	open domain
	TransOMCS Zhang *et al.* [2020a]	
	CSKG Ilievski *et al.* [2021]	
	ATOMIC Sap *et al.* [2019a]	human interaction
	ATOMIC$^{20}_{20}$ Hwang *et al.* [2021]	
	ASER Zhang *et al.* [2020b]	eventuality

Table 1: Common knowledge sources used in PLMKEs.

2.3 Characteristics of Current Knowledge Sources

We discuss two typical characteristics of current knowledge sources: **large-scale** and **diverse**. Regarding to the scale of knowledge sources, all the encyclopedic knowledge sources in Table 1 contain millions of concepts and at least hundred million of facts induced by them. The largest commonsense sources in Table 1 is ASER, which contains 64 million facts. We observe that the size of commonsense knowledge sources is much smaller than the encyclopedic ones. However, compared with prior commonsense sources such as Cyc [Lenat, 1995], the current sources are produced in more precise and scalable way: the annotation process is partially automatic and accessible to non-experts. The current trend of commonsense collection methods manifests the potential to scale up the knowledge sources.

The domains that common knowledge sources cover are diverse. Encyclopedic knowledge sources such as Wikipedia, DBPedia and Freebase are collected from open domain, suggesting that they are constructed by heterogeneous knowledge not limited to specific domains. Meanwhile, knowledge sources involving specific domains such as biomedicine and science (e.g., UMLS and AMiner) are established to boost

the development of domain-specific applications. Since commonsense involves various aspects including human interaction and object properties in everyday life, there exist both open-domain commonsense knowledge sources (e.g., ConceptNet, TransOMCS) that cover multiple domains of commonsense, and domain-specific commonsense sources (e.g., ATOMIC, ASER) that focus on particular types. The diversity of knowledge sources is beneficial for laying the foundations of broader future applications of PLMKEs.

3 Knowledge-Intensive NLP Tasks

Knowledge-intensive NLP tasks are served as testbeds to evaluate the capability of PLMKEs to solve problems that require external knowledge. In this section, we provide an overview of knowledge-intensive NLP tasks, and summarize their corresponding features.

3.1 Overview of Knowledge-Intensive NLP Tasks

Knowledge-intensive NLP tasks can be divided into two groups based on the types of the required knowledge: encyclopedic and commonsense knowledge-intensive tasks.

Table 2 lists several typical datasets of three representative **encyclopedic knowledge-intensive tasks**: open-domain question answering (QA), fact verification, and entity linking. The open-domain QA task aims to answer information seeking questions (e.g., "When was Barack Obama born?") that require encyclopedic knowledge over various domains. Fact verification (e.g., judge if the claim "Barack Obama born was born on August 4 1961." is true) is designed to verify whether a given text claim is factually correct which also demands the model's ability to reason over large amount of factual knowledge. Lastly, the purpose of entity linking system is to link text mentions of entities to their corresponding unique identifier in a target database, e.g., linking to Wikipedia pages. The datasets [Thorne *et al.*, 2018; Kwiatkowski *et al.*, 2019; Guo and Barbosa, 2014] for these tasks heavily relies on encyclopedic knowledge sources including Wikipedia and DBPedia.

Commonsense knowledge-intensive tasks focus on testing whether models can accurately understand and respond to daily scenarios. For example, in a task about social commonsense, given a context like "Someone spilled the food all over the floor", models are required to select the most proper responses like "He/she needs to mop up" instead of unreasonable ones like "Run around in the mess". The types of commonsense-intensive tasks are diverse because of the diversity of commonsense knowledge. As shown in Table 3, tasks about social commonsense involve human interactions and thoughts; when it comes to physical commonsense, the questions in the datasets inquire physical properties and ways to manipulate objects; temporal commonsense reasoning datasets usually contain questions about event order, duration and frequency.

3.2 Characteristics of Knowledge-Intensive Tasks

The most salient characteristic is that external encyclopedic or commonsense knowledge is necessary to perform well on these tasks. Not only for models, it is hard for humans to answer questions about Barack Obama's birth date without any reference knowledge. Another characteristic is that although these tasks can be tackled with external knowledge, the required knowledge may not be directly given along with input. In other words, the default task input is simply a question or context of a certain scenario, without any additional information. It motivates researchers to consider adding a module in charge of grounding to external knowledge sources in the design of PLMKEs.

[1]https://lemurproject.org/clueweb12/

Tasks	Datasets	Data Sources
Open-domain QA	NATURAL QUESTIONS Kwiatkowski *et al.* [2019]	Wikipedia
	HOTPOTQA Yang *et al.* [2018]	Wikipedia
Fact Verification	FEVER Thorne *et al.* [2018]	Wikipedia
	BOOLQ Clark *et al.* [2019]	Wikipedia
Entity Linking	ACE2004 NIST [2004]	news
	AIDA CONLL-YAGO Hoffart *et al.* [2011]	DBPedia & YAGO Suchanek *et al.* [2007]
	WNWI Guo and Barbosa [2014]	Wikipedia
	WNCW Guo and Barbosa [2014]	Clueweb [1]

Table 2: Detailed information about representative encyclopedic knowledge-intensive tasks and datasets.

Commonsense Types	Tasks/Datasets	Data Sources
General Commonsense	COMMONSENSEQA Talmor *et al.* [2019]	ConceptNet
	WSC Levesque *et al.* [2012]	human thoughts
	αNLI Bhagavatula *et al.* [2020]	ROCStories Mostafazadeh *et al.* [2016]
	COMMONGEN Lin *et al.* [2020]	image captions & ConceptNet
Social Commonsense	SOCIALIQA Sap *et al.* [2019b]	ATOMIC Sap *et al.* [2019a]
Physical Commonsense	PIQA Bisk *et al.* [2020]	image captions & ConceptNet
	HELLASWAG Zellers *et al.* [2019]	video captions
Temporal Commonsense	MCTACO Zhou *et al.* [2019a]	MultiRC Khashabi *et al.* [2018]
	TACRIE Zhou *et al.* [2021]	ROCStories Mostafazadeh *et al.* [2016]

Table 3: Detailed information about types of commonsense and representative commonsense knowledge-intensive tasks.

4 Knowledge Fusion Methods

Tackling knowledge-intensive NLP tasks highly relies on the assistance of appropriate knowledge sources. How to fuse such important knowledge into the strong pre-trained language models and make them more knowledgeable poses challenges to researchers.

The training process of pre-trained language models usu-

ally involve two stages: pre-training and fine-tuning [Devlin *et al.*, 2019]. Pre-training is a self-supervised learning process to learn representations through language modeling on large unlabeled text corpus. Fine-tuning is the process to adapt pre-trained models with task-specific supervision on downstream tasks. In PLMKEs, knowledge can be integrated in either stage but the fusion methods for the two stages are quite different. In this section, we categorize the mainstream knowledge fusion methods by the stage where knowledge is fused in PLMKEs.

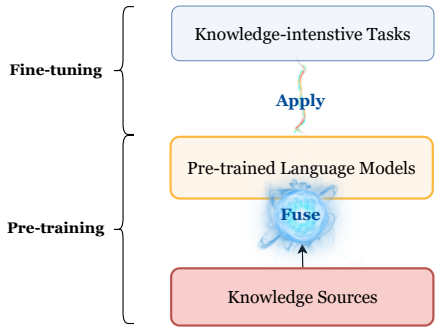

Figure 1: Pre-fusion methods.

4.1 Pre-Fusion Methods

Pre-fusion methods fuse external knowledge in the **pre-training stage** [Zhang *et al.*, 2019; Sun *et al.*, 2021]. Before knowledge fusion, knowledge sources are first processed into the format similar to unstructured raw text corpus. Then, the processed text corpus are used for further pre-training with a sharing set of learning objectives used by original language models. Thus, pre-fusion methods enable knowledge fusion without much architectural change. For knowledge sources such as knowledge graphs, however, the knowledge is usually structured and not aligned with the unstructured input format of language models. The simplest approach to tackling this challenge is concatenating the entities and relation [Zhang *et al.*, 2019] or generating fluent synthetic sentences by conditional text generation models [Agarwal *et al.*, 2021].

4.2 Post-Fusion Methods

Instead of incorporating knowledge in the pre-training stage, post-fusion methods seek to fuse the knowledge in the **fine-tuning stage** [Zhou *et al.*, 2019b; Lin *et al.*, 2019; Liu *et al.*, 2020; Cheng *et al.*, 2021]. Given an input text from a particular knowledge-intensive task, post-fusion methods first retrieve the relevant knowledge to the input, and then perform a joint reasoning on top of the augmented input.

To capture relevant knowledge, the post-fusion methods leverage text retriever on unstructured knowledge sources to extract the useful textual spans; for structured knowledge sources, previous works attempt to match the entities appearing in input text to the relevant entity-centric subgraphs. After capturing the knowledge, they select one of the two following approaches to implementing knowledge fusion. The captured knowledge can be transformed into knowledge embeddings by encoders such as graph neural networks, and used as supplemental features to the text embeddings provided by pre-training language models for the following reasoning modules (structure-aware post-fusion: path A in Figure 2) [Lin *et al.*, 2019; Yasunaga *et al.*, 2021]. It can also be directly concatenated with the text input and fed into the pre-training language models altogether (text-based post-fusion: path B in Figure 2) [Karpukhin *et al.*, 2020; Cheng *et al.*, 2021].

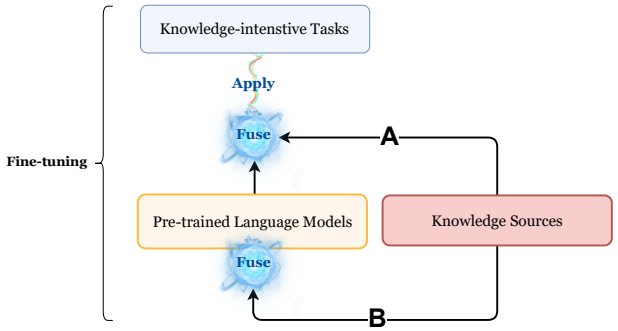

Figure 2: Post-fusion methods. Path A indicates structure-aware post-fusion methods that encode structured knowledge into supplemental embeddings and integrate with the text embeddings produced by pre-trained language models. Path B indicates text-based post-fusion methods that retrieve relevant texts and subgraphs and convert them into texts to be fed into pre-trained language models. Commonly, each PLMKE chooses either path A or B to fuse knowledge.

4.3 Hybrid-Fusion Methods

Hybrid-fusion methods are a combination of pre-fusion and post-fusion methods: knowledge is fused in both **pre-training** and **fine-tuning stages**. Although there exists salient difference between pre-fusion and post-fusion methods, the hybrid-fusion methods enable us to unify both: the retriever frequently leveraged in post-fusion methods can be jointly trained in the pre-training stage. That is, during the pre-training, language models are also learning to leverage additional retrieved knowledge for modeling the language context. The pre-trained model augmented by the jointly learned retriever can thus utilize the knowledge from the retriever more effectively during the fine-tuning stage. The retrieval-augmented pre-training [Peters *et al.*, 2019; Guu *et al.*, 2020; Lewis *et al.*, 2020] is commonly adopted in hybrid-fusion methods and it manifests the effectiveness on several knowledge-intensive tasks discussed in Section 3.

4.4 Representative Models for Specific Tasks

Here, we first categorize representative models for various knowledge-intensive NLP tasks based on their corresponding knowledge fusion methods discussed in previous section. Table 4 and 5 list state-of-the-art (SOTA) PLMKEs on encyclopedic and commonsense knowledge-intensive tasks and several representative models, respectively. For encyclopedic knowledge-intensive tasks, it is shown that except BoolQ, the other state-of-the-art models all adopt post-fusion methods. On the contrary, for commonsense knowledge-intensive tasks, except CommonsenseQA, pre-fusion methods are broadly leveraged in the SOTA models.

Tasks	Datasets	Models	Fusion Types	Fused Knowledge
Open-domain QA	NATURAL QUESTIONS Kwiatkowski *et al.* [2019]	UnitedQA Cheng *et al.* [2021]	Post-fusion	Wikipedia
	WEBQUESTION Berant *et al.* [2013]	EMDR^2 Sachan *et al.* [2021]	Post-fusion	Wikipedia
	TRIVIAQA Joshi *et al.* [2017]	EMDR^2 Sachan *et al.* [2021]	Post-fusion	Wikipedia
	Other Representative Models	REALM Guu *et al.* [2020]	Hybrid-fusion	Wikipedia
		RAG Lewis *et al.* [2020]	Hybrid-fusion	Wikipedia
Fact Verification	FEVER Thorne *et al.* [2018]	Jiang *et al.* [2021]	Post-fusion	Wikipedia
	BOOLQ Berant *et al.* [2013]	ERNIE 3.0 Sun *et al.* [2021]	Pre-fusion	Wikipedia, Bookcorpus Zhu *et al.* [2015], etc.
	Other Representative Models	GEAR Zhou *et al.* [2019b]	Post-fusion	Wikipedia
		KGAT Liu *et al.* [2020]	Post-fusion	Wikipedia
Entity Linking	AIDA CONLL-YAGO Hoffart *et al.* [2011]	Mulang' *et al.* [2020]	Post-fusion	Wikipedia
	Other Representative Models	CHOLAN Ravi *et al.* [2021]	Post-fusion	Wikipedia

Table 4: State-of-the-art PLMKEs on encyclopedic knowledge-intensive tasks and other representative models.

Commonsense Types	Tasks/Datasets	Models	Fusion Types	Fused Knowledge
General Commonsense	COMMONSENSEQA Talmor *et al.* [2019]	GreaseLM Zhang *et al.* [2022]	Post-fusion	ConceptNet
	WSC Levesque *et al.* [2012]	ERNIE 3.0 Sun *et al.* [2021]	Pre-fusion	Wikipedia, Bookcorpus Zhu *et al.* [2015], etc.
	αNLI Bhagavatula *et al.* [2020]	UNIMO Li *et al.* [2021]	Pre-fusion	Wikipedia, Bookcorpus, images
	Other Representative Models	KagNet Lin *et al.* [2019]	Post-fusion	ConceptNet
		QA-GNN Yasunaga *et al.* [2021]	Post-fusion	ConceptNet
Social Commonsense	SOCIALIQA Sap *et al.* [2019b]	UNICORN Lourie *et al.* [2021]	Pre-fusion	various commonsense benchmarks
	Other Representative Models	UnifiedQA-11B Khashabi *et al.* [2020]	Pre-fusion	various QA benchmarks
		McQueen Mitra *et al.* [2019]	Post-fusion	ATOMIC
Physical Commonsense	PIQA Bisk *et al.* [2020]	UNICORN Lourie *et al.* [2021]	Pre-fusion	various commonsense benchmarks
	Other Representative Models	UnifiedQA-11B Khashabi *et al.* [2020]	Pre-fusion	various QA benchmarks

Table 5: State-of-the-art PLMKEs on commonsense knowledge-intensive tasks and other representative models.

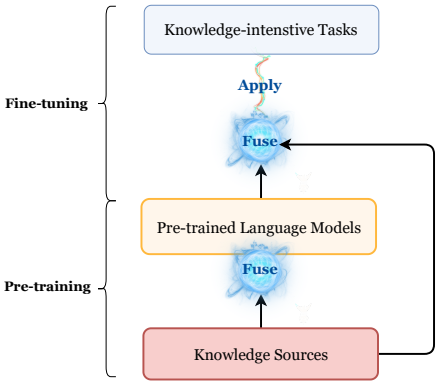

Figure 3: Hybrid-fusion methods.

We then analyze why pre-fusion methods are not prevalent and effective on encyclopedic knowledge-intensive tasks. In pre-fusion methods, the knowledge required in these tasks is implicitly stored in pre-trained parameters. But it is hard to determine what the knowledge is finally stored, and it also increases the difficulty in eliciting and leveraging the knowledge [Guu *et al.*, 2020; Wang *et al.*, 2021]. Instead, post-fusion methods are able to infer upon explicit and concrete textual knowledge. But the advantage of post-fusion to leverage explicit and concrete knowledge may become a shortcoming for commonsense knowledge-intensive tasks. As mentioned in Section 2, commonsense is usually implicit inside texts and the coverage of commonsense knowledge sources is much smaller than that of encyclopedic knowledge sources. Even if we compose large-scale commonsense knowledge sources with the help of knowledge acquisition methods, we are still likely to miss a large body of commonsense knowledge used in our daily life. Therefore, it is possible that post-fusion methods may fail at retrieving the relevant knowledge inside the knowledge sources and thus cannot bring extra benefits to the pre-trained language models.

5 Challenges and Future Directions

5.1 Unified PLMKEs Across Tasks and Domains

Recent developments of PLMKEs have led to task-specific modeling advances. As shown in Table 4, the models frequently used on encyclopedic knowledge-intensive tasks adopt post-fusion and hybrid-fusion methods, while the two fusion methods are not usually exploited on commonsense knowledge-intensive tasks. Furthermore, we observe that the state-of-the-art models in different knowledge-intensive NLP tasks are unique, making the progress on each task seemingly incompatible. Beyond the tasks listed in Table 2 and 3, knowledge-intensive NLP tasks are extended to various domains involving biomedical and legal [Liu *et al.*, 2021b] knowledge. Recently, researchers also attach more importance to the diversity of knowledge existing in different times [Dhingra *et al.*, 2021] and regions [Zhang and Choi, 2021; Yin *et al.*, 2021; Liu *et al.*, 2021a]. The diversity across tasks and domains is naturally fostering the need for unified PLMKEs, instead of promoting the trend of devising unique models on individual tasks.

5.2 Reliabilty of Knowledge Sources

Since knowledge sources are the basis of PLMKEs, we are concerned with the reliability of knowledge sources. Currently, many large-scale knowledge sources are constructed by automatic knowledge acquisition algorithms. Whereas, there exists a trade-off between the scale and precision: it is likely to introduce false and biased information into the knowledge sources [Sun and Peng, 2021]. We anticipate bias amplification in case the PLMKEs are constructed on biased knowledge sources. We call upon researchers to be aware of the reliability of knowledge sources via proposing more precise knowledge acquisition algorithms and careful inspection over the knowledge sources they intend to use.

5.3 Reasoning Module Design

Reasoning is an important step for solving knowledge-intensive NLP tasks. It is essential especially for commonsense knowledge-intensive tasks, since the relevant commonsense knowledge is usually implicit and should be used in multiple turns of reasoning. When we humans encounter the situation like "Someone spilled the food all over the floor", we are first aware of the fact that the floor is not clean, and others' shoes would get dirty if they stepped on the spilled food. Based on the situation, the intent to mop up the floor produces. Though several PLMKEs achieve great performance on such commonsense-related scenarios, it is unclear about whether the models perform human-like reasoning implicitly or simply capture the spurious correlation and thus not robust upon more complex situations. Existing models containing reasoning modules mainly perform reasoning on entity [Lin *et al.*, 2019; Yasunaga *et al.*, 2021] or syntax structures [Yin *et al.*, 2020; Bai *et al.*, 2021], which cannot cover the complex situations like the aforementioned "spill food" example. To achieve the human-like capability of recognizing the everyday situations, multi-hop reasoning module is needed for designing trustworthy PLMKEs that can simulate human thoughts.

6 Conclusions

We comprehensively survey existing works about knowledge-intensive NLP with pre-trained language models and summarize the current progress in terms of the three critical components in PLMKEs: knowledge sources, knowledge-intensive NLP tasks, and knowledge fusion methods. Based on the discussion about the three components, we further pose several challenges that would be influential in the practical usage and propose the related future directions in response to the challenges. We hope that this paper could provide NLP practitioners with a clear picture on the topic and boost the development of the current knowledge-intensive NLP technologies.

References

Oshin Agarwal, Heming Ge, et al. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. In *ACL*, 2021.

Sören Auer, Christian Bizer, et al. DBPedia: A Nucleus for a Web of Open Data. In *The Semantic Web*. 2007.

Jiangang Bai, Yujing Wang, et al. Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees. In *EACL*, 2021.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic Parsing on Freebase from Question-Answer Pairs. In *EMNLP*, 2013.

Chandra Bhagavatula, Ronan Le Bras, et al. Abductive Commonsense Reasoning. In *ICLR*, 2020.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*, 2020.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 2004.

Kurt Bollacker, Colin Evans, et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *ICDM*, 2008.

Hao Cheng, Yelong Shen, et al. UnitedQA: A Hybrid Approach for Open Domain Question Answering. In *ACL*, 2021.

Christopher Clark, Kenton Lee, et al. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *ACL*, 2019.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Knowledge Neurons in Pretrained Transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 2019.

Bhuwan Dhingra, Jeremy R Cole, et al. Time-Aware Language Models as Temporal Knowledge Bases. *arXiv preprint arXiv:2106.15110*, 2021.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*, 2022.

Zhaochen Guo and Denilson Barbosa. Robust Entity Linking via Random Walks. In *CIKM*, 2014.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. In *ICML*, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-Enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*, 2020.

Johannes Hoffart, Mohamed Amir Yosef, et al. Robust Disambiguation of Named Entities in Text. In *EMNLP*, 2011.

Jena D. Hwang, Chandra Bhagavatula, et al. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *AAAI*, 2021.

Filip Ilievski, Pedro A. Szekely, and Bin Zhang. CSKG: The CommonSense Knowledge Graph. In *ESWC*, 2021.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. Exploring Listwise Evidence Reasoning with T5 for Fact Verification. In *ACL*, 2021.

Mandar Joshi, Eunsol Choi, et al. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*, 2017.

Vladimir Karpukhin, Barlas Oguz, et al. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*, 2020.

Daniel Khashabi, Snigdha Chaturvedi, et al. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *ACL*, 2018.

Daniel Khashabi, Sewon Min, et al. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *ACL*, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, et al. Natural Questions: A Benchmark for Question Answering Research. *TACL*, 2019.

Douglas B Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 1995.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *KR*, 2012.

Patrick S. H. Lewis, Ethan Perez, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*, 2020.

Wei Li, Can Gao, et al. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *ACL*, 2021.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP*, 2019.

Bill Yuchen Lin, Wangchunshu Zhou, et al. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *ACL*, 2020.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained Fact Verification with Kernel Graph Attention Network. In *ACL*, 2020.

Fangyu Liu, Emanuele Bugliarello, et al. Visually Grounded Reasoning across Languages and Cultures. In *EMNLP*, 2021.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. Everything Has a Cause: Leveraging Causal Inference in Legal Text Analysis. In *NAACL*, 2021.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on Rainbow: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. *arXiv preprint arXiv:2103.13009*, 2021.

Christopher D. Manning, Kevin Clark, et al. Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-Supervision. *PNAS*, 2020.

Arindam Mitra, Pratyay Banerjee, et al. How Additional Knowledge can Improve Natural Language Commonsense Question Answering? *arXiv preprint arXiv:1909.08855*, 2019.

Nasrin Mostafazadeh, Nathanael Chambers, et al. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *ACL*, 2016.

Isaiah Onando Mulang', Kuldeep Singh, et al. Evaluating the Impact of Knowledge Graph Context on Entity Disambiguation Models. In *CIKM*, 2020.

US NIST. The Ace Evaluation Plan. *US National Institute for Standards and Technology (NIST)*, 2004.

Matthew E Peters, Mark Neumann, et al. Knowledge Enhanced Contextual Word Representations. In *EMNLP*, 2019.

Fabio Petroni, Tim Rocktäschel, et al. Language Models as Knowledge Bases? In *EMNLP*, 2019.

Fabio Petroni, Aleksandra Piktus, et al. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *ACL*, 2021.

Alec Radford, Jeff Wu, et al. Language Models are Unsupervised Multitask Learners. 2019.

Manoj Prabhakar Kannan Ravi, Kuldeep Singh, et al. CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata. In *EACL*, 2021.

Adam Roberts, Colin Raffel, and Noam Shazeer. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Prof. of EMNLP*, 2020.

Devendra Singh Sachan, Siva Reddy, et al. End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering. *CoRR*, 2021.

Maarten Sap, Ronan Le Bras, et al. ATOMIC: An Atlas of Machine Commonsense for If-then Reasoning. In *AAAI*, 2019.

Maarten Sap, Hannah Rashkin, et al. Social IQa: Commonsense Reasoning about Social Interactions. In *EMNLP*, 2019.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 2017.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge. In *WWW*, 2007.

Jiao Sun and Nanyun Peng. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *NAACL*, 2021.

Yu Sun, Shuohuan Wang, et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *CoRR*, 2021.

Alon Talmor, Jonathan Herzig, et al. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *ACL*, 2019.

Jie Tang, Jing Zhang, et al. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD*, 2008.

James Thorne, Andreas Vlachos, et al. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *ACL*, 2018.

Denny Vrandečić. Wikidata: A New Platform for Collaborative Data Collection. In *WWW*, 2012.

Alex Wang, Amanpreet Singh, et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. 2018.

Alex Wang, Yada Pruksachatkun, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *ICONIP*, 2019.

Cunxiang Wang, Pai Liu, and Yue Zhang. Can Generative Pre-trained Language Models Serve As Knowledge Bases for Closed-book QA? In *ACL*, 2021.

Xiaokai Wei, Shen Wang, et al. Knowledge Enhanced Pretrained Language Models: A Compreshensive Survey. *arXiv preprint arXiv:2110.08455*, 2021.

Zhilin Yang, Peng Qi, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*, 2018.

Jian Yang, Gang Xiao, et al. A Survey of Knowledge Enhanced Pre-trained Models. *arXiv preprint arXiv:2110.00269*, 2021.

Michihiro Yasunaga, Hongyu Ren, et al. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *ACL*, 2021.

Da Yin, Tao Meng, and Kai-Wei Chang. SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics. In *ACL*, 2020.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning. In *EMNLP*, 2021.

Rowan Zellers, Ari Holtzman, et al. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*, 2019.

Michael Zhang and Eunsol Choi. SituatedQA: Incorporating Extra-Linguistic Contexts into QA. In *EMNLP*, 2021.

Zhengyan Zhang, Xu Han, et al. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*, 2019.

Hongming Zhang, Daniel Khashabi, et al. TransOMCS: From Linguistic Graphs to Commonsense Knowledge. In *IJCAI*, 2020.

Hongming Zhang, Xin Liu, et al. ASER: A Large-scale Eventuality Knowledge Graph. In *WWW*, 2020.

Xikun Zhang, Antoine Bosselut, et al. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. *arXiv preprint arXiv:2201.08860*, 2022.

Ben Zhou, Daniel Khashabi, et al. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *EMNLP*, 2019.

Jie Zhou, Xu Han, et al. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *ACL*, 2019.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal Reasoning on Implicit Events from Distant Supervision. In *ACL*, 2021.

Yukun Zhu, Ryan Kiros, et al. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *ICCV*, 2015.