Relational World Knowledge Representation in Contextual Language Models: A Review

Tara Safavi, Danai Koutra
University of Michigan, Ann Arbor
{tsafavi,dkoutra}@umich.edu

Abstract

Relational **knowledge bases** (KBs) are commonly used to represent world knowledge in machines. However, while advantageous for their high degree of precision and interpretability, KBs are usually organized according to manually-defined schemas, which limit their expressiveness and require significant human efforts to engineer and maintain. In this review, we take a natural language processing perspective to these limitations, examining how they may be addressed in part by training deep contextual **language models** (LMs) to internalize and express relational knowledge in more flexible forms. We propose to organize knowledge representation strategies in LMs by the level of KB supervision provided, from no KB supervision at all to entity- and relation-level supervision. Our contributions are threefold: **(1)** We provide a high-level, extensible taxonomy for knowledge representation in LMs; **(2)** Within our taxonomy, we highlight notable models, evaluation tasks, and findings, in order to provide an up-to-date review of current knowledge representation capabilities in LMs; and **(3)** We suggest future research directions that build upon the complementary aspects of LMs and KBs as knowledge representations.

1 Introduction

Knowledge bases (**KBs**) are data structures that connect pairs of entities or concepts by semantically meaningful symbolic relations. Decades' worth of research have been invested into using KBs as tools for relational world knowledge representation in machines (Minsky, 1974; Lenat, 1995; Liu and Singh, 2004; Bollacker et al., 2008; Vrandečić and Krötzsch, 2014; Speer et al., 2017; Sap et al., 2019; Ilievski et al., 2021).

Most large-scale modern KBs are organized according to a manually engineered schema that specifies which entity and relation types are permitted, and how such types may interact with one another. This explicit enforcement of relational structure is both an advantage and a drawback (Halevy et al., 2003). On one hand, schemas support complex queries over the data with accurate, consistent, and interpretable answers. On the other hand, schemas are "ontological commitments" (Davis et al., 1993) that limit flexibility in how knowledge is stored, expressed, and accessed. Handcrafted schemas also require significant human engineering effort to construct and maintain, and are therefore often highly incomplete (Weikum et al., 2021).

Language models as KBs? The tension between structured and unstructured knowledge representations is not new in natural language processing (Banko and Etzioni, 2008; Fader et al., 2011). However, only recently has an especially promising solution emerged, brought about by breakthroughs in machine learning software, hardware, and data. Specifically, deep contextual language models (**LMs**) like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) have shown to be capable of internalizing a degree of relational world knowledge within their parameters, and expressing this knowledge across various mediums and tasks—in some cases, *without* the need for a predefined entity-relation schema (Petroni et al., 2019; Roberts et al., 2020). Consequently, some have begun to wonder whether LMs will partially or even fully replace KBs, given sufficiently large training budgets and parameter capacities.

Present work In this review, we summarize recent compelling progress in machine representation of relational world knowledge with LMs. We propose to organize relevant work by the level of KB supervision provided to the LM (Figure 1):

- **Word-level supervision** (§ 3): At this level, LMs are not explicitly supervised on a KB, but may be indirectly exposed to KB-like knowledge via word associations in the training corpus. Here, we cover techniques for probing

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1053–1067
November 7–11, 2021.

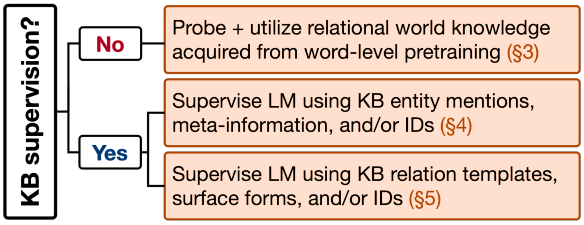


Figure 1: A high-level overview of our taxonomy, organized by level of KB supervision provided.

and utilizing this implicitly acquired knowledge.

- **Entity-level supervision** (§ 4): At this level, LMs are supervised to acquire knowledge of KB entities. Here, we organize strategies from "less symbolic" to "more symbolic": Less symbolic approaches train LMs with entity-aware language modeling losses, but never explicitly require the LM to link entity mentions to the KB. By contrast, more symbolic approaches involve linking, and may also integrate entity embeddings into the LM's parameters.
- **Relation-level supervision** (§ 5): At this level, LMs are supervised to acquire knowledge of KB triples and paths. Again, we organize strategies from less to more symbolic, where less symbolic approaches treat triples as fully natural language statements, and more symbolic approaches incorporate dedicated embeddings of KB relation types.

For each supervision level, we provide notable examples in terms of methodology and/or findings, and compare the benefits and drawbacks of different approaches. We conclude in § 6 with our vision of the future, emphasizing the complementary roles of LMs and KBs as knowledge representations.

Related work As this topic is relatively nascent, few related surveys exist. Closest to our own work, Colon-Hernandez et al. (2021) cover methods for combining contextual language representations with graph representations, albeit with a comparatively narrow scope and no discussion of implicit knowledge. Liu et al. (2021a) survey prompt-based learning in LMs, which overlaps with our discussion of cloze prompting in § 3.1, although relational world knowledge is not their main focus.

2 Preliminaries

We briefly review preliminaries and assumptions necessary for our survey.

Knowledge bases We use the term "knowledge base" (**KB**) to refer to a relational data structure comprising a set of **entities** E, **relation types** R, and **triples** $(s, r, o) \in E \times R \times E$, where $s, o \in E$ are subject and object entities, respectively.[1] We consider two types of KBs under the umbrella of "relational world knowledge." **Encyclopedic** KBs store facts about typed, disambiguated entities; a well-known example is the Wikidata KB (Vrandečić and Krötzsch, 2014), which, like its sister project Wikipedia, is publicly accessible and collaboratively constructed. By contrast, in **commonsense** KBs, "entities" are typically represented by non-canonicalized free-text phrases. Examples include the publicly accessible, crowdsourced ConceptNet (Liu and Singh, 2004; Speer et al., 2017) and ATOMIC (Sap et al., 2019) KBs.

Language models Following the contemporary NLP literature, we use the term "language model" (**LM**) to refer to a deep neural network that is trained to learn contextual text representations. LMs generally come **pretrained**, with parameters pre-initialized for generic text representation via self-supervised training on large corpora, and may be used as-is after pretraining, or further **fine-tuned** with supervision on downstream task(s). This work considers LMs based on the **Transformer** architecture (Vaswani et al., 2017), examples of which include the encoder-only BERT family (Devlin et al., 2019; Liu et al., 2019), the decoder-only GPT family (Brown et al., 2020), and the encoder-decoder T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) families.

3 Word-level supervision

The standard language modeling task is to predict the n-th word in a sequence of n words—that is, a conditional probability estimation task (Radford et al., 2019). While many variants of this task have been proposed to allow LMs to condition their predictions on different inputs (Devlin et al., 2019; Raffel et al., 2020; Lewis et al., 2020), a notable feature of all such approaches is that they operate at the word (and subword) level.

If these supervision techniques do not incorporate KBs at all, how are they relevant when considering LMs as relational knowledge representations? The answer is simple. Typical language

[1] For our purposes, we consider the terms "knowledge base" and "knowledge graph" as interchangeable.

Table 1: Taxonomy and representative examples for extracting relational knowledge in word-level pretrained LMs, with evaluation tasks that have been conducted in the referenced papers. *Glossary of evaluation tasks*: KP—knowledge probing; QA—question answering; CR—compositional reasoning; KC—knowledge base construction.

Knowledge extracted via...	Extraction strategy	Representative examples	Evaluation task(s)			
			KP	QA	CR	KC
Cloze prompts (§ 3.1)	Prompt handcrafting	(Petroni et al., 2019; Dufter et al., 2021)	✓			
	Automatic prompt engineering	(Jiang et al., 2020b; Shin et al., 2020; Zhong et al., 2021; Qin and Eisner, 2021)	✓			
	Adversarial prompt modification	(Kassner and Schütze, 2020; Petroni et al., 2020; Poerner et al., 2020; Cao et al., 2021)	✓			
	Varying base prompts	(Elazar et al., 2021; Heinzerling and Inui, 2021; Jiang et al., 2020a; Kassner et al., 2021)	✓			
	Symbolic rule-based prompting	(Kassner et al., 2020; Talmor et al., 2020a)	✓		✓	
Statement scores (§ 3.2)	Single-LM scoring	(Tamborrino et al., 2020; Zhou et al., 2020)		✓	✓	
	Dual-LM scoring	(Davison et al., 2019; Shwartz et al., 2020)		✓		✓

modeling corpora like Wikipedia are known to contain KB-like assertions about the world (Da and Kasai, 2019). LMs trained on enough such data can be expected to acquire some KB-like knowledge, even without targeted entity- or relation-level supervision. Therefore, in order to motivate the necessity (if at all) of KB supervision, it is crucial to first understand what relational world "knowledge" LMs acquire from word-level pretraining. In this section, we cover strategies to extract and utilize this knowledge under the cloze prompting (§ 3.1) and statement scoring (§ 3.2) protocols. Table 1 provides a taxonomy for this section, with representative examples and evaluation tasks.

3.1 Cloze prompting

The cloze prompting protocol (Taylor, 1953 and Figure 2) is a direct approach for extracting and evaluating KB-like knowledge in pretrained LMs. Under this protocol, KB triples are first converted to natural language assertions using (e.g.) relation templates. For each assertion, the token(s) corresponding to the object entity are held out. A frozen pretrained LM then ranks candidate tokens within its vocabulary by the probability that they fill in the empty slot(s). Accuracy is typically measured by the proportion of prompts for which the correct answer appears in the LM's top-k predictions, with the assumption that better performance implies more pretrained knowledge within the LM.

Handcrafted prompts in English with single-token answers make up LAMA (Petroni et al., 2019), one of the earliest and most widely-used LM cloze probes. LAMA, which is mapped primarily to Wikidata and ConceptNet triples, was initially used to compare pretrained LMs' knowledge to off-the-shelf KB question answering systems. Petroni et al. (2019) showed that pretrained BERT is competitive with a supervised relation extraction model that has been provided an oracle for entity linking, particularly for 1-1 queries. Subsequent work has experimented with handcrafted templates for probing the knowledge of both very large (hundred-billion parameter) LMs (Brown et al., 2020) as well as non-contextual word embeddings, i.e., as a simple control baseline for LMs (Dufter et al., 2021). Both studies demonstrate some success, particularly in cases where the probed model is provided a small amount of extra context in the form of conditioning examples (Brown et al., 2020) or entity type information (Dufter et al., 2021).

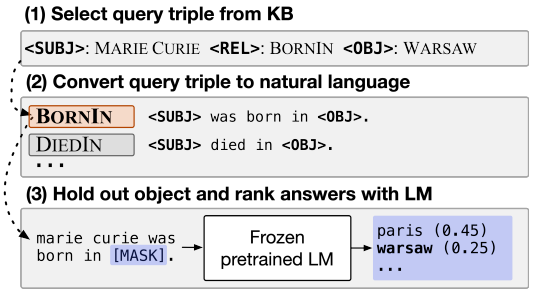


Figure 2: Probing relational knowledge in pretrained LMs with cloze prompts generated from KB triples.

Automatic prompt engineering is a promising alternative to prompt handcrafting for knowledge extraction in LMs (Liu et al., 2021a), as prompts engineered using discrete (Jiang et al., 2020b; Shin et al., 2020; Haviv et al., 2021) and continuous (Zhong et al., 2021; Qin and Eisner, 2021; Liu et al., 2021b) optimization have improved LMs' lower-bound performance on LAMA's underlying queries. Note, however, that optimized prompts are not always grammatical or intelligible (Shin et al., 2020). Prompt optimization methods may also confound knowledge probes by overfitting to the probes' answer distributions during train-

ing (Zhong et al., 2021; Cao et al., 2021), and often require large validation sets for tuning, which may not be feasible in practice (Perez et al., 2021).

Adversarial modification of LAMA prompts has uncovered weaknesses in pretrained LMs' world "knowledge," for example that BERT's accuracy drops precipitously when irrelevant statements or negation words are added to prompts (Kassner and Schütze, 2020; Lin et al., 2020; Petroni et al., 2020), and that it can "guess" answers using shallow lexical cues or benchmark artifacts (Poerner et al., 2020; Cao et al., 2021). However, the adversarial robustness of LM knowledge improves greatly with supervision in both the pretraining (Petroni et al., 2020) and fine-tuning (Kassner and Schütze, 2020) stages, suggesting that explicit KB-level supervision is a viable remedy to input sensitivity.

Several collections of prompt variations, including paraphrased sets of base prompts (Elazar et al., 2021; Heinzerling and Inui, 2021) and multilingual sets of base (English) prompts (Jiang et al., 2020a; Kassner et al., 2021) have been released to expand the original research questions posed by LAMA. For the former, it has been found that pretrained BERT-based LMs typically do not output consistent answers for prompt paraphrases, although their consistency can again be greatly improved by targeted pretraining (Elazar et al., 2021; Heinzerling and Inui, 2021). For the latter, initial results on prompts beyond English indicate high variability in pretrained LM performance across languages and poor performance on prompts with multi-token answers (Jiang et al., 2020a; Kassner et al., 2021).

Prompts generated with symbolic rules have been used to test pretrained LMs' abilities to learn, e.g., equivalence, implication, composition, and conjunction. Existing studies vary the degrees of experimental control: Talmor et al. (2020a) use BERT-based models with their publicly-available pretrained weights, whereas Kassner et al. (2020) pretrain BERT from scratch on synthetic KB triples only. Both studies observe mixed results, concluding that word-level pretraining alone (at least on BERT) does not lead to strong "reasoning" skills.

3.2 Statement scoring

Beyond probing, pretrained LM "knowledge" can be purposed toward downstream KB-level tasks in a zero-shot manner via statement scoring. Here, a pretrained LM is fed natural language statements corresponding to KB triples, and its token probabilities across each statement are pooled to yield statement scores. These scores are then treated as input to a downstream decision, mirroring the way that supervised LMs can be trained to output probabilities for triple-level prediction tasks (§ 5). We categorize statement scoring strategies as single- or dual-LM approaches. The **single-LM** approach pools the pretrained LM's token scores over a candidate set of sequences, then takes the highest-scoring sequence as the LM's "prediction" or choice (Tamborrino et al., 2020; Bouraoui et al., 2020; Zhou et al., 2020; Brown et al., 2020). The **dual-LM** framework first uses one pretrained LM to generate useful context (e.g., clarification text) for the task, then feeds this context to another, possibly different pretrained LM to obtain a final score (Davison et al., 2019; Shwartz et al., 2020).

Both categories have shown promise over comparable unsupervised (and, under some conditions, supervised) methods for tasks like multiple-choice QA (Tamborrino et al., 2020; Shwartz et al., 2020; Brown et al., 2020) and commonsense KB completion (Davison et al., 2019). However, LM scores have also shown to be sensitive to small perturbations in text (Zhou et al., 2020), so this approach may be less effective on noisy or long-tail inputs.

3.3 Summary and outlook

There is still broad disagreement over the nature of acquired "knowledge" in pretrained LMs. Whereas some studies suggest that word-level pretraining may be enough to endow LMs with KB-like knowledge (Petroni et al., 2019; Tamborrino et al., 2020), in particular given enough parameters and the right set of prompts (Brown et al., 2020), others conclude that such pretraining alone does not yield sufficiently precise or robust LM knowledge (Elazar et al., 2021; Cao et al., 2021)—directly motivating the targeted supervision strategies discussed in the remainder of this paper. We observe that different studies independently set objectives for what a pretrained LM should "know," and thus naturally reach different conclusions. We believe that future studies must reach consensus on standardized tasks and benchmarks, addressing questions like: What degree of overlap between a pretraining corpus and a knowledge probe is permissible, and how can this be accurately uncovered and quantified? What lexical cues or correlations should be allowed in knowledge probes? Progress in this direction will

Table 2: Taxonomy and representative examples of entity-level supervision in LMs, with evaluation tasks that have been conducted in the referenced papers. *Glossary of evaluation tasks*: KP—knowledge probing; EL—entity linking; ET—entity typing; RC—relation classification; QA—question answering; GL—the General Language Understanding Evaluation or GLUE benchmark (Wang et al., 2019), which covers multiple subtasks.

Entities as...	Supervision strategy	Representative examples	Evaluation task(s)					
			KP	EL	ET	RC	QA	GL
Token mention-spans (§ 4.1)	Masked token prediction	(Roberts et al., 2020; Guu et al., 2020)					✓	
	Contrastive learning	(Xiong et al., 2020; Shen et al., 2020)	✓		✓		✓	
Text-to-KB links—late fusion (§ 4.2)	Linking w/o external info	(Broscheit, 2019; Ling et al., 2020)		✓				✓
	Linking w/ textual metadata	(Wu et al., 2020; De Cao et al., 2021)		✓		✓	✓	
	Linking w/ external embeddings	(Zhang et al., 2019; Chen et al., 2020)		✓	✓	✓		✓
Text-to-KB links—mid/early fusion (§ 4.3)	Entity embedding retrieval	(Peters et al., 2019; Févry et al., 2020)	✓	✓	✓	✓	✓	
	Treating entities as tokens	(Yamada et al., 2020; Poerner et al., 2020)	✓	✓	✓	✓	✓	

not only further our understanding of the effects of word-level supervision on LM knowledge acquisition, but will also provide appropriate yardsticks for measuring the benefits of targeted entity- and relation-level supervision.

4 Entity-level supervision

We next review entity-level supervision strategies for LMs, most often toward improving performance in knowledge probes like LAMA (§ 3.1) and canonical NLP tasks like entity typing, entity linking, and question answering. We roughly categorize approaches from "least symbolic" to "most symbolic." On the former end of the spectrum, the LM is exposed to entity mentions in text but not required to link these mentions to an external entity bank (§ 4.1). On the latter end, the LM is trained to link mentions to the KB using late (§ 4.2) or mid-to-early fusion approaches (§ 4.3). Table 2 provides a taxonomy of supervision strategies for this section with representative examples.

4.1 Modeling entities without linking

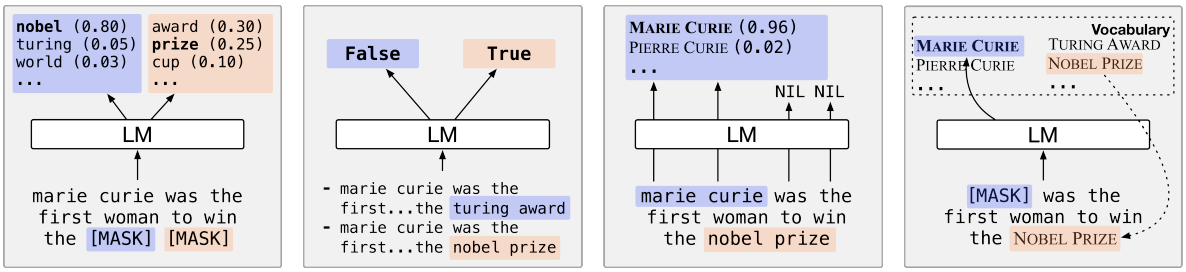
The "least symbolic" entity supervision approaches that we consider input textual contexts containing entity mention-spans to the LM, and incorporate these mention-spans into their losses. However, they do not require the LM to link these mentions to the KB's entity set, so the LM is never directly exposed to the KB. Figures 3a and 3b provide examples of input and output for this class of approaches.

Masking tokens in mention-spans and training LMs to predict these tokens may promote knowledge memorization (Sun et al., 2020). Roberts et al. (2020) investigate this strategy using a simple masking strategy whereby an LM is trained to predict the tokens comprising named entities and dates in text (Figure 3a, originally proposed by Guu et al., 2020). The authors find that the largest (11 billion parameter) version of T5 generates exact-match answers on open-domain question answering (QA) benchmarks with higher accuracy than extractive systems—even without access to external context documents, simulating a "closed-book" exam.

Contrastive learning techniques, which have been used for LM supervision at the word and sentence level (Devlin et al., 2019), have also been devised for supervision on entity mentions (Shen et al., 2020). For example, Xiong et al. (2020) replace a proportion of entity mentions in the pretraining corpus with the names of negatively-sampled entities of the same type, and train an LM to predict whether the entity in the span has been replaced (Figure 3b). Although the previously discussed closed-book T5 model (Roberts et al., 2020) outperforms Xiong et al. (2020)'s open-book BERT pretrained with contrastive entity replacement on open-domain QA, the latter may generalize better: T5's performance degrades considerably for facts not observed during training, whereas open-book approaches appear more robust (Lewis et al., 2021).

4.2 Linking with late fusion

The next-strongest level of entity supervision is to train the LM to link entity-centric textual contexts to a KB's entity set E. Here, we cover late fusion approaches, which operate at the word level in terms of input to the LM and incorporate entities at the LM's output layer only, as exemplified in Figure 3c. The simplest representatives of this category train LMs to match individual tokens (Broscheit, 2019) or mentions (Ling et al., 2020) in a text corpus to an entity bank, without any external resources. The minimally "entity-aware" BERT proposed by Broscheit (2019), which adds a single classification layer on top of a pretrained BERT encoder, achieves competitive results with a state-of-the-art specialized entity linking architec-



(a) Mention-span masking (b) Contrastive learning (c) Linking—late fusion (d) Linking—early fusion

Figure 3: Examples of entity-level supervision in LMs, ranging from "less symbolic" to "more symbolic."

ture (Kolitsas et al., 2018).

Entity meta-information such as names and descriptions are viable external resources for LM-powered entity linking (Botha et al., 2020). For example, in zero-shot entity linking (Logeswaran et al., 2019), textual mentions must be linked to entities unseen during training using only entity descriptions as additional data. Here, competitive solutions train separate BERT models to select and rank candidate entities by encoding their descriptions (Logeswaran et al., 2019; Wu et al., 2020). More recently, encoder-decoder LMs have been trained to retrieve entities by generating their unique names (De Cao et al., 2021), which has the advantage of scaling with the LM's vocabulary size (usually tens of thousands) instead of the KB entity set size (potentially tens of millions). De Cao et al. (2021) achieve results competitive to discriminative approaches on entity linking and QA, suggesting the potential of generative entity-aware LMs.

External entity embeddings pretrained by a separate model have been used as strong sources of inductive bias for LMs. For example, several variants of BERT further pretrain the base model by linearly fusing external entity embeddings with contextual word representations at the output of the BERT encoder (Zhang et al., 2019; He et al., 2020). BERT has also been fine-tuned to match its output token representations to external entity embeddings for the task of end-to-end entity linking (Chen et al., 2020). Such approaches rely heavily on the quality of the externally-learned embeddings, which is both a strength and a drawback: Such embeddings may contain useful implicit structural information about the KB, but on the other hand may propagate errors into the LM (Shen et al., 2020).

4.3 Linking with middle or early fusion

The last and strongest category of entity supervision techniques that we consider are also linking-based, but fuse entity information at earlier stages of text encoding. Mid-fusion approaches retrieve external entity representations in between hidden layers and re-contextualize them into the LM, whereas early fusion approaches simply treat entity symbols as tokens in the vocabulary. Figure 3d provides an example of input/output for early fusion.

Retrieving entity embeddings and integrating them into an LM's hidden word representations is a middle-fusion technique that has the advantage of modeling flexibility: It allows the practitioner to choose where (i.e., at which layer) the entity embeddings are integrated, and how the entity embeddings are learned and re-contextualized into the LM. Peters et al. (2019) integrate externally pre-trained, frozen entity embeddings into BERT's final hidden layers using a word-to-entity attention mechanism. Févry et al. (2020) learn the external entity embeddings jointly during pretraining, and perform the integration in BERT's earlier hidden layers using an attention-weighted sum. The latter approach is competitive with a 30× larger T5 LM in closed-book QA (§ 4.1), suggesting that LMs and KB embeddings can be trained jointly to enhance and complement each other.

Treating entities as "tokens" by appending special reserved entity symbols to the LM's vocabulary is the earliest of entity fusion approaches (Figure 3d). For instance, Yamada et al. (2020) input entity "tokens" alongside textual contexts that mention these entities to RoBERTa, and use specialized word-to-entity and entity-to-entity attention matrices within its hidden layers. Other approaches leave the base LM's internal architecture completely unchanged and focus only on aligning the LM's word and entity embedding spaces at the input level (Rosset et al., 2020; Poerner et al., 2020). Note, however, that this approach may significantly enlarge the LM's vocabulary. For example, plain BERT's vocabulary is around 30k tokens, whereas

Table 3: Taxonomy and representative examples of relation-level supervision in LMs, with evaluation tasks conducted in the respective referenced papers. *Glossary of evaluation tasks*: KP—knowledge probing; ET—entity typing; RC—relation classification; QA—question answering; CR—compositional reasoning; KC—knowledge base construction; TG—text generation; GL—the GLUE family of language tasks (Wang et al., 2019).

Relations as...	Supervision strategy	Representative examples	Evaluation task(s)							
			KP	ET	RC	QA	CR	KC	TG	GL
Templated sentences (§ 5.1)	Lexicalizing triples	(Thorne et al., 2021; Guan et al., 2020)				✓	✓		✓	
	Lexicalizing paths	(Clark et al., 2020; Talmor et al., 2020a,b)	✓				✓			
Linearized sequences (§ 5.2)	Training on triple sequences	(Yao et al., 2019; Agarwal et al., 2021)	✓			✓		✓	✓	
	Injecting triples into text	(Liu et al., 2020)				✓				
Dedicated embeddings (§ 5.3)	Pooling entity representations	(Baldini Soares et al., 2019; Qin et al., 2021)		✓	✓	✓				
	Embedding relations externally	(Wang et al., 2021d; Daza et al., 2021)		✓	✓			✓		✓
	Treating relations as tokens	(Bosselut et al., 2019; Hwang et al., 2021)						✓		

English Wikipedia has around 6 million entities. This can make pretraining on a larger vocabulary expensive in terms of both time and memory usage (Yamada et al., 2020; Dufter et al., 2021).

4.4 Summary and outlook

The literature on entity supervision in LMs is growing rapidly. In line with recent trends in NLP (Khashabi et al., 2020), a growing number of entity supervision strategies use generative models (Roberts et al., 2020; De Cao et al., 2021), which are attractive because they allow for a high level of flexibility in output and circumvent the need for classification over potentially millions of entities. However, some studies find that generative models currently do not perform well beyond what they have memorized from the training set (Wang et al., 2021b; Lewis et al., 2021). These findings suggest that storing some entity knowledge externally (e.g., in a dense memory, Févry et al., 2020) may be more robust, for example by allowing for efficient updates to the LM's knowledge (Verga et al., 2020). We believe that future work will need to analyze the tradeoffs between fully-parametric and retrieval-based entity modeling in terms of pure accuracy, parameter and training efficiency, and ability to generalize beyond the training set.

5 Relation-level supervision

Finally, we consider methods that utilize KB triples or paths to supervise LMs for complex, often compositional tasks like relation classification, text generation, and rule-based inference. We again organize methods in the order of less to more symbolic. In this context, less symbolic approaches treat triples and paths as fully natural language (§ 5.1, 5.2). By contrast, more symbolic approaches learn distinct embeddings for relation types in the KB (§ 5.3). Table 3 provides a taxonomy of this section with representative examples and evaluation tasks.

5.1 Relations as templated assertions

Template-based lexicalization is a popular relation supervision strategy that does not directly expose the LM to the KB. Similar to how KB queries are converted to cloze prompts for knowledge probing (§ 3.1), triples are first converted to natural language assertions using relation templates, usually handcrafted. These assertions are then fed as input to the LM, which is trained with any number of task-specific losses. Figure 4 provides an input/output example for this class of approach.

Lexicalized triples from Wikidata have been used as LM training data in proof-of-concept studies demonstrating that LMs can serve as natural language querying interfaces to KBs under controlled conditions (Heinzerling and Inui, 2021). A promising approach in this direction uses encoder-decoder LMs to generate answer sets to natural language queries over lexicalized Wikidata triples (Thorne et al., 2020, 2021), toward handling multi-answer KB queries with LMs—thus far an understudied task in the LM knowledge querying literature.

Other approaches convert KB triples to sentences using relation templates in order to construct task-specific training datasets for improved performance in, e.g., story generation (Guan et al., 2020), commonsense QA (Ye et al., 2020; Ma et al., 2021), and relation classification (Bouraoui et al., 2020). While most of these approaches rely on template handcrafting, a few automatically mine templates using distant supervision on Wikipedia, achieving competitive results in tasks like relation classification (Bouraoui et al., 2020) and commonsense QA (Ye et al., 2020).

Compositional paths spanning multiple atoms of symbolic knowledge may also be lexicalized and input to an LM (Lauscher et al., 2020; Talmor et al., 2020a) in order to train LMs for soft com-

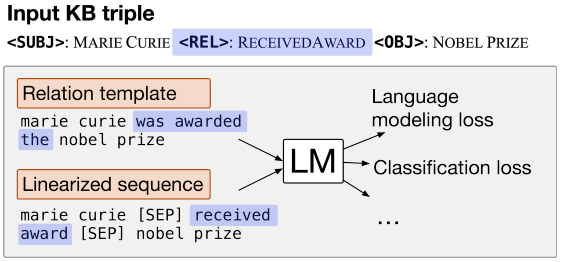


Figure 4: Strategies for representing relations as sequences: Templating (§ 5.1) and linearization (§ 5.2).

positional reasoning (Clark et al., 2020; Talmor et al., 2020b). Notably, when RoBERTa is fine-tuned on sentences expressing (real or synthetic) facts and rules from a KB, it can answer entailment queries with high accuracy (Clark et al., 2020; Talmor et al., 2020b). However, as Clark et al. (2020) note, these results do not necessarily confirm that LMs can "reason," but rather that they can at least emulate soft reasoning—raising an open question about how to develop probes and metrics to verify whether LMs can actually reason compositionally.

5.2 Linearizing KB triples

The main advantage of templating is that it converts symbolic triples into sequences, which can be straightforwardly input to LMs. However, handcrafting templates is a manual process, and distant supervision can be noisy. To maintain the advantage of templates while avoiding the drawbacks, triples can alternatively be fed to an LM by linearizing them—that is, flattening the subject, relation, and object into an input sequence (Figure 4). With linearization, relation-level supervision becomes as simple as **feeding the linearized sequences** to the LM and training again with task-specific losses (Yao et al., 2019; Kim et al., 2020; Ribeiro et al., 2020; Wang et al., 2021a) or **injecting the sequences into the pretraining corpus** (Liu et al., 2020). A notable recent example of the former approach (Agarwal et al., 2021) trains T5 on linearized Wikidata triples in order to generate fully natural language versions of those triples. These verbalized triples are used as retrieval "documents" for improved LM-based QA over traditional document corpora; note, however, that they can also be used as LM training data for other downstream tasks in place of handcrafted templates (§ 5.1).

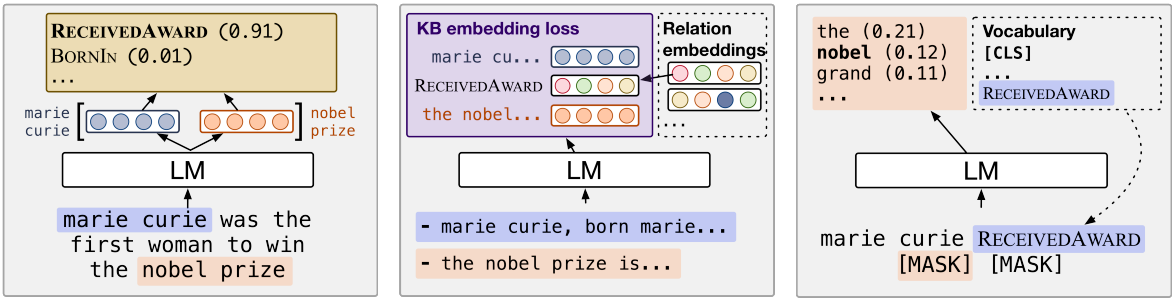
5.3 Relations as dedicated embeddings

The strategies discussed thus far treat KB triples and paths as natural language sequences. A "more symbolic" approach is to represent KB relation types with dedicated embeddings, and integrate these embeddings into the LM using late, middle, or early fusion approaches. Figures 5a and 5b provide input/output examples for late fusion, whereby relation textual contexts are input to the LM, and relation embeddings are constructed or integrated at the LM's output. Figure 5c exemplifies early fusion, whereby relations are treated as input tokens.

Contextual representations of entity mention-spans may be pooled at an LM's output layer to represent a relation (Wang et al., 2021c; Yu et al., 2020). For example, Baldini Soares et al. (2019) concatenate the contextual representations of special entity-start markers inserted adjacent to textual entity mentions, and fine-tune BERT to output similar relation representations for statements ranging over the same entity pairs (Figure 5a). This approach, which proved highly successful for relation classification, has been applied to the same task in languages beyond English (Köksal and Özgür, 2020; Ananthram et al., 2020), and as an additional LM pretraining objective (Qin et al., 2021).

Non-contextual relation embeddings may be learned by defining a separate relation embedding matrix with $|R|$ rows and fusing this matrix into the LM. One advantage of this approach, similar to methods for retrieving external entity embeddings (§ 4.3), is that it supports fusion at both the late (Wang et al., 2021d; Daza et al., 2021) and middle (Liu et al., 2021c) stages. As an example of the former, Wang et al. (2021d) propose an LM pretraining objective whereby textual descriptions of KB entities are input to and encoded by an LM, then combined with externally-learned relation embeddings at the output using a link prediction loss (Figure 5b). Combined with standard word-level language modeling objectives, this approach enables generalization across both sentence-level tasks like relation classification, and graph-level tasks like KB completion.

Treating relations as "tokens," toward early fusion of relations in LMs, is achieved by appending the KB's relation types to the LM's vocabulary (Figure 5c). A notable instantiation of this approach is the COMET commonsense KB construction framework (Bosselut et al., 2019; Hwang et al., 2021; Jiang et al., 2021). Given a subject phrase/relation token as input, COMET fine-tunes an LM to generate object phrases. COMET demon-



(a) Late fusion—pooling (b) Late fusion—external embeddings (c) Early fusion—relations as "tokens"

Figure 5: Examples of relation supervision strategies that incorporate dedicated embeddings of relation types.

strates promising improvements over 400× larger LMs not trained for KB construction (Hwang et al., 2021). However, templating (§ 5.1) may yield better results than adding special tokens to the vocabulary when the COMET framework is trained and tested in a few-shot setting (Da et al., 2021).

5.4 Summary and outlook

Relation-level supervision in LMs is exciting because it enables a wide variety of complex NLP tasks (Table 3). A unifying theme across many of these tasks is that of compositionality, or the idea that smaller "building blocks" of evidence can be combined to arrive at novel knowledge. As compositionality is thought to be key to machine generalization (Lake et al., 2017), we believe that further fundamental research in understanding and improving LMs' soft "reasoning" skills (Clark et al., 2020; Talmor et al., 2020b, § 5.1) will be crucial.

Finally, while most of the open directions we discuss involve improving LM knowledge with KBs, we find the direction of generating KBs with LMs equally intriguing—reflecting the fact that LMs and KBs can complement each other in "both directions," as automating and scaling out the construction of KBs will ultimately provide LMs with more relational training data. The generative COMET framework (Bosselut et al., 2019, § 5.3) has made inroads in commonsense KB construction, but the same progress has not yet been observed for encyclopedic knowledge. The latter entails unique challenges: Whereas commonsense entities are not disambiguated and triples need only be plausible rather than always true, encyclopedic entities are usually disambiguated and facts are often binary true/false. We look forward to future research that addresses these challenges, perhaps building on recent breakthroughs in generative factual entity retrieval (De Cao et al., 2021, § 4.2).

6 Conclusion and vision

In this review, we provide an overview of how LMs may acquire relational world knowledge during pretraining and fine-tuning. We propose a novel taxonomy that classifies knowledge representation methodologies based on the level of KB supervision provided to an LM, from no explicit supervision at all to entity- and relation-level supervision.

In the future, we envision a stronger synergy between the perspectives and tools from the language modeling and knowledge bases communities. In particular, we expect powerful and expressive LMs, which are actively being developed in NLP, to be increasingly combined with large-scale KB resources to improve their knowledge recall and reasoning abilities. On the converse, we expect such KB resources to be increasingly generated directly by LMs. Within both of these directions, we hope that future work will continue to explore the themes discussed in this paper, in particular that of delineating and testing KB-level memorization versus generalization in LMs. We also expect that more standardized benchmarks and tasks for evaluating LM knowledge will be developed, a direction that has recently seen some progress (Petroni et al., 2021). As research at the intersection of LMs and KBs is rapidly progressing, we look forward to new research that better develops and combines the strengths of both knowledge representations.

Acknowledgements

We thank the reviewers for their thoughtful feedback. This material is based upon work supported by the National Science Foundation under a Graduate Research Fellowship and CAREER Grant No. IIS 1845491, the Advanced Machine Learning Collaborative Grant from Procter & Gamble, and an Amazon faculty award.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Large scale knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*.

Amith Ananthram, Emily Allaway, and Kathleen McKeown. 2020. Event-guided denoising for multilingual relation learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1505–1512, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463.

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 pre-proceedings*.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. 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)*, pages 1860–1874, Online. Association for Computational Linguistics.

Haotian Chen, Xi Li, Andrej Zukov Gregoric, and Sahil Wadhwa. 2020. Contextualized end-to-end neural entity linking. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 637–642, Suzhou, China. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3882–3890.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*.

Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, and Antoine Bosselut. 2021. Understanding few-shot commonsense knowledge models. *arXiv preprint arXiv:2101.00297*.

Jeff Da and Jungo Kasai. 2019. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 1–12, Hong Kong, China. Association for Computational Linguistics.

Randall Davis, Howard Shrobe, and Peter Szolovits. 1993. What is a knowledge representation? *AI magazine*, 14(1):17–17.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. 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)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Daniel Daza, Michael Cochez, and Paul Groth. 2021. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*, pages 798–808.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter, Nora Kassner, and Hinrich Schütze. 2021. Static embeddings as efficient knowledge bases? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *arXiv preprint arXiv:2102.01017*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Alon Y Halevy, Oren Etzioni, AnHai Doan, Zachary G Ives, Jayant Madhavan, Luke K McDowell, and Igor Tatarinov. 2003. Crossing the structure chasm. In *Conference on Innovative Data Systems Research*.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online. Association for Computational Linguistics.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290, Online. Association for Computational Linguistics.

Benjamin Heinzerling and Kentaro Inui. 2021. 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*, pages 1772–1791, Online. Association for Computational Linguistics.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. *European Semantic Web Conference*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "i'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. In *The 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. 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*, pages 7811–7818, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1737–1743, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 340–350, Online. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. *arXiv preprint arXiv:2001.03765*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. 2021c. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Marvin Minsky. 1974. A framework for representing knowledge.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint arXiv:2105.11447*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Knowledge-aware language model pretraining. *arXiv preprint arXiv:2007.00655*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8980–8994, Online. 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020a. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020b. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *34th Conference on Neural Information Processing Systems*.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2020. From natural language processing to neural databases. *Proceedings of the VLDB Endowment*, 14(6):1033–1039.

James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. Database reasoning over text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3091–3104, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William W Cohen. 2020. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. *arXiv preprint arXiv:2007.00849*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 2021*, pages 1737–1748.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021b. Can generative pre-trained language models serve as knowledge bases for closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251, Online. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021c. K-adapter: Infusing knowledge into pre-trained models with adapters. In *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. 2021d. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguistics*, 9:176–194.

Gerhard Weikum, Xin Luna Dong, Simon Razniewski, and Fabian M. Suchanek. 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. *Found. Trends Databases*, 10(2-4):108–490.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2020. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint arXiv:1908.06725*.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020. Jaket: Joint pre-training of knowledge graph and language understanding. *arXiv preprint arXiv:2010.00796*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. 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*, pages 5017–5033, Online. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.