Improving Language Model Predictions via Prompts Enriched with Knowledge Graphs*

Ryan Brate[1], Minh-Hoang Dang[2], Fabian Hoppe[3,4], Yuan He[5], Albert Meroño-Peñuela[6] and Vijay Sadashivaiah[7]

[1]*KNAW Humanities Cluster, Digital Humanities Lab, Amsterdam, Netherlands*
[2]*LS2N, Université de Nantes, Faculté des Sciences et Techniques (FST), France*
[3]*FIZ Karlsruhe, Leibniz Institute for Information Infrastructure, Germany*
[4]*Karlsruhe Institute of Technology, Institute AIFB, Germany*
[5]*University of Oxford, UK*
[6]*King's College London, UK*
[7]*Rensselaer Polytechnic Institute, USA*

Abstract

Despite advances in deep learning and knowledge graphs (KGs), using language models for natural language understanding and question answering remains a challenging task. Pre-trained language models (PLMs) have shown to be able to leverage contextual information, to complete cloze prompts, next sentence completion and question answering tasks in various domains. Unlike structured data querying in e.g. KGs, mapping an input question to data that may or may not be stored by the language model is not a simple task. Recent studies have highlighted the improvements that can be made to the quality of information retrieved from PLMs by adding auxiliary data to otherwise naive prompts. In this paper, we explore the effects of enriching prompts with additional contextual information leveraged from the Wikidata KG on language model performance. Specifically, we compare the performance of naive vs. KG-engineered cloze prompts for entity genre classification in the movie domain. Selecting a broad range of commonly available Wikidata properties, we show that enrichment of cloze-style prompts with Wikidata information can result in a significantly higher recall for the investigated BERT and RoBERTa large PLMs. However, it is also apparent that the optimum level of data enrichment differs between models.

Keywords

Prompt Learning, Pre-trained Language Model, Knowledge Graph.

1. Introduction

Pre-trained language models (PLMs) [1, 2], based on deep learning attention-based architectures, have shown to have outstanding performance at various natural language processing (NLP) tasks predicated on natural language understanding. However, the extent to which they capture domain knowledge and *empirical semantics* [3] — i.e. the use of formal domain properties

Workshop on Deep Learning for Knowledge Graphs (DL4KG@ISWC2022), October 23-24, 2022
* Authors listed in alphabetical order
r.brate@gmail.com (R. Brate); minhhoangdang@hotmail.com (M. Dang); fabian.hoppe@kit.edu (F. Hoppe); yuan.he@cs.ox.ac.uk (Y. He); albert.merono@kcl.ac.uk (A. Meroño-Peñuela); sadasv2@rpi.edu (V. Sadashivaiah)
0000-0002-7047-2770 (F. Hoppe); 0000-0002-4486-1262 (Y. He); 0000-0003-3375-3810 (V. Sadashivaiah)

in practice — is not well understood. In this work, we narrow down the focus to cloze-style completion, the task of predicting the masked entity text in a sentence. For example, given: "The Klingons are a species in the franchise [MASK]", the PLM is expected to predict "Star Trek" for [MASK]. It aims to extract the implicit knowledge entailed by the PLMs, since such knowledge can be used for downstream NLP applications like sentiment analysis [4], dialogue systems [5], and natural language inference [6], as well as for completing the missing information of knowledge graphs (KGs) or ontologies [7], and even constructing new ones [8].

In recent years, PLMs have improved on the state of the art in many NLP tasks by leveraging large text corpora [9], but most of time they still require annotated data for task-specific fine-tuning [10]. However, the empirical semantics gathered by these models is limited to distributional aspects [11]. Therefore, the performance, especially in the few- and zero-shot setting, highly depends on the provided *prompt*, i.e. snippets of contextual information for a specific task. However, in many cases the engineering of the prompts is naive and simplistic, giving the PLM too little context to provide an accurate answer, and unsystematic, providing little principles on how exactly these prompts need to be composed in order to have a predictable behaviour. Indeed, recent studies [12] have highlighted the improvements that can be made to the quality of information retrieved from PLMs by performing amendments to these prompts. This casts doubts on some studies [13] that claim that a PLM cannot answer easy questions about e.g. culture (movies, books, music, ...), it is reasonable to postulate that PLMs could perhaps answer those questions accurately if they were provided with systematically engineered prompts that contained richer contexts.

Existing approaches of prompt engineering include: *(i)* learn-by-example, where the prompt consists of the concatenation of correct examples we expect a PLM to predict [2]; *(ii)* manually designed prompts of different granularities [13]; *(iii)* automatically searched prompts optimized on few-shot samples [14], all of which rely on the implicit semantics of natural language texts. In this paper, we investigate how incorporating explicit knowledge from external sources like KGs can help prompt engineering and thus enhance the cloze-style question answering of PLMs. Specifically, we explore cloze-style prompts with respect to the movie domain in respect of the performance of the BERT and RoBERTa large PLMs.

2. Related Work

Studies towards prompt learning are based on the hypothesis that pre-trained language models (PLMs) have learnt abundant knowledge and just require sufficiently detailed contexts for predictions [2, 10, 15] — and in this way, it is possible to apply PLMs without data-driven fine-tuning. A (hard[1]) *prompt* is the conditioning text which is combinded with the input to provide contexts or hints for the PLM. A *template* (i.e. *pattern*) is a function that integrates the inputs and prompts. Answers are then given by the PLMs conditioned on the prompts, and a further function (i.e. *verbalizer*) is often required to map the answers to the final outputs. The reason for that is, the prompt learning paradigm is typically formulated as a similar task to the PLM's pre-training task, which does not necessarily yield the desired outputs of downstream applications.

[1] *Soft* prompts are learnt at the embedding level.

An important part of prompt learning is prompt engineering, i.e., to design template(s), either manually or automatically, to support downstream applications. In [2], Brown et al. proposed to use *demonstrations*, i.e., a sequence of input-output texts, as the prompts, expecting that the PLM can implicitly learn to predict from examples. For instance, if we want the PLM to predict the masked position in "[MASK] is the capital of China.", we can demonstrate by appending "London is the capital of the UK" after the masked sentence. Schick et al. [16] manually designed different templates, each corresponding to an individual PLM trained on few-shot examples. The predictions of downstream text classification and natural language inference tasks were then made according to an ensemble of trained PLMs. Shin et al. [14] argued that manually designed templates suffer from the uncertainty of guesswork or the lack of domain expertise. Therefore, they proposed to search for templates using gradient-based optimization. More recently, Lu et al. [17] have shown that PLMs performance varies with the order of these prompts, and use generative language models and entropy statistics on the prompt permutations to identify prompts with good performance.

KGs or ontologies are excellent sources for providing explicit knowledge to enrich prompts or verbalizers. West et al. [18] considered *distilling* a student model in the common sense domain from the enormously large PLM GPT-3 [2], which serves as the teacher model. They adopted the prompt learning scheme to extract triples from the teacher model with templates created and examples extracted from the common sense KG Atomic [19]. Hu et al. [7] argued that the label word space (i.e., the answer space) can be well expanded by adding in external knowledge about related words. They employed different refinement heuristics to shortlist candidates to benefit the downstream classification task. For instance, if some "Person" is classified as a "Physicist" in the ground truth data, then answers like "Scientist" will also be accepted.

Our work was motivated by the probing study of Penha et al. [13] that investigates whether BERT (a well-known PLM consisting of stacked transformer encoders [1]) actually knows superficial cultural knowledge about books, movies, and music. Cloze-style questions for classifying the genre of entities (from Wikidata) of different books, movies, and music were given for the PLM to answer, often with unsatisfying performance. However, their work considered naive prompts without sufficient contexts, while ours attempts to examine if KGs can enrich these prompts, especially giving additional contexts (e.g., attributes, k-hop neighbours) of the entities in order to help the PLM to generate better predictions.

3. Methodology

The basic idea of our method is to use the information about entities in KGs to expand cloze-style prompts with richer entity descriptions. It is summarized in Figure 1. We enrich the naive prompt, for example `Die Hard is of genre [MASK]`, through matching the movie *Die Hard* to the corresponding Wikidata item and extract auxiliary knowledge with SPARQL queries, and generating an enriched prompt using this auxiliary data. We use datatype properties and verbalize entities using *rdfs:label* to compose valid phrases. As a result, we obtain e.g. `Die Hard is a movie, starring Bruce Willis, directed by John McTier- nan, of the genre [MASK].`

We then use both (a) the naive prompts and (b) the KG-enriched prompts to query various

language models, and compare their performance on the entity genre classification task. In the following paragraphs the enrichment by KG querying and the prompt engineering step are described in detail.

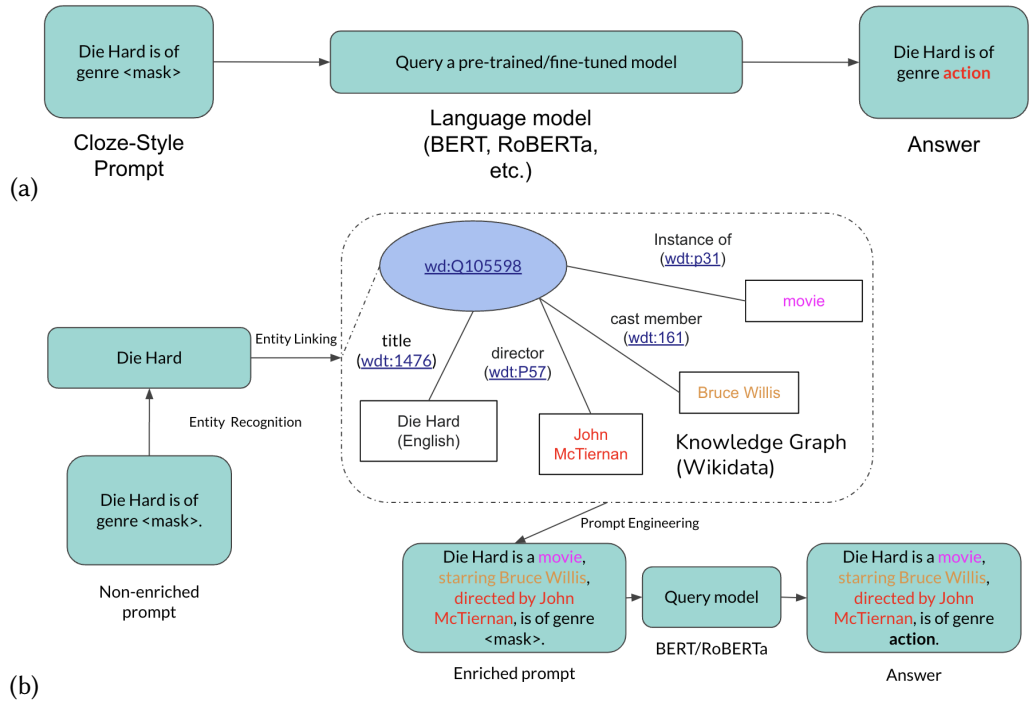

Figure 1: Proposed framework (a) typical querying setup for a Masked Language Model prediction. (b) Proposed approach to enrich the query using external language.

3.1. Knowledge Graph Querying

The auxiliary data for each movie is extracted from Wikidata. This is done in a simplistic two-step-process using SPARQL queries. The queries operate on a batch of input records to reduce the number of requests and avoid timeout errors.

First, the movies are linked to their respected Wikidata entities by IMDb or TMDB ID utilizing the Wikidata properties *IMDb ID* (wdt:P345) and *TMDb movie ID* (wdt:P4947). If this does not provide an entity, an exact string matching given the title is attempted as well.

```
SELECT ?mlId ?imdbId ?tmdbId ?movie
WHERE {
   VALUES (?mlId ?imdbId ?tmdbId) {("1" "tt0114709" "862" ) ... }
   {?movie wdt:P345 ?imdbId . }
   UNION
   {?movie wdt:P4947 ?tmdbId .}
}
```

Listing 1: SPARQL query used for entity linking with the IMDb or TMDB ID.

The second step queries the entities for the auxiliary data used to enrich the prompts with

additional contextual information. Overall, a set of 28 properties was extracted and investigated for each entity. A simplified version of the utilized SPARQL query is given in 2. This query can easily be adapted to query other properties by adding these properties to the *?property* values. From this set of properties a subset of 10 manually selected domain-specific properties are used to constract the enriched prompts. The properties are selected based on human intuition and the most frequent co-occurrence for the given entities.

```
SELECT ?mlId (SAMPLE(?movieLabel) AS ?movieLabel) (SAMPLE(?propertyLabel) AS ?propertyLabel)
      (GROUP_CONCAT(DISTINCT ?objectLabel; SEPARATOR=", ") AS ?objectList)
WHERE{
   VALUES (?mlId ?movie) { ("1" ) ...}
   ?movie rdfs:label ?movieLabel .
   FILTER (LANG(?movieLabel)="en")
   VALUES ?property {wdt:P144 wdt:P179 ...}

   ?p1 wikibase:directClaim ?property .
   ?p1 rdfs:label ?propertyLabel .
   FILTER (LANG(?propertyLabel)="en")
   OPTIONAL {
      ?movie ?property ?object .
      ?object rdfs:label ?objectLabel .
      FILTER (LANG(?objectLabel)="en")
   }
}
GROUP BY ?mlId ?property
```

Listing 2: Simplified SPARQL query used to retrieve additional movie knowledge from Wikidata.

3.2. Prompt Engineering

Similarly to [13], we consider an entity genre classification task. The prompts are of the form: "<title> is a movie <Wikidata enrichment>, of the genre [MASK].", where <Wikidata enrichment> is an aggregation of movie properties and corresponding values extracted from Wikidata pertaining the title in question, in some natural language format. Table 1 lists the Wikidata properties used to assemble values for <Wikidata enrichment>.

Wikidata property	Property Label	Enrichment Text
wdt:P161	cast member	starring
wdt:P57	director	directed by
wdt:P162	producer	produced by
wdt:P58	screenwriter	screenwriter
wdt:P86	composer	music by
wdt:P1040	film editor	edited by
wdt:P577	year	released
wdt:P750	distributed by	distributed by
wdt:P495	country of origin	originating from

Table 1
Wikidata properties used in constructing probes for the movie dataset. '*Enrichment Text*' is the text adopted in the probe enrichment to describe the property in question in a more natural language format.

The Wikidata properties listed in Table 1 are broadly ranked in descending information

specificity. It was in this order, that ten variations for a probe were constructed, by sequentially adding Wikidata properties to prompts, building gradually more contextual-information dense prompts. In adding property information, only the first value of each Wikidata property was used where more than one was available (e.g., the first listed cast member). E.g., as follows; the unenriched prompt, the first two successive prompt enrichments, and the final enriched form pertaining to the movie Die Hard.:

- non-enriched prompt: Die Hard is a movie, of the genre [MASK].
- enriched Prompt 1(A): Die Hard is a movie, *starring Bruce Willis,* of the genre [MASK].
- enriched Prompt 2(A): Die Hard is a movie, *starring Bruce Willis, directed by John McTiernan,* of the genre [MASK].
- enriched Prompt 9(A): Die Hard is a movie, *starring Bruce Willis, directed by John McTiernan, produced by Joel Silver, screenwriter Roderick Thorp, music by Michael Kamen, edited by John F. Link, released 1988, distributed by Netflix, originating from United States of America,* of the genre [MASK].

Given the potential for sensitivity of PLMs to the verbalisation strategy used in the construction of cloze-stype prompts, we considered two verbalisation strategies for aggregation of the additional Wikidata properties. Whereas the above *verbalisation strategy A* form is aggregated with commas, the *verbalisation strategy B* form is aggregated with *and* tokens. E.g.:

- enriched Prompt 1(B): Die Hard is a movie *and starring Bruce Willis,* of the genre [MASK].
- enriched Prompt 2(B): Die Hard is a movie *and starring Bruce Willis and directed by John McTiernan,* of the genre [MASK].
- enriched Prompt 9(B): Die Hard is a movie *and starring Bruce Willis and directed by John McTiernan and produced by Joel Silver and screenwriter Roderick Thorp and music by Michael Kamen and edited by John F. Link and released 1988 and distributed by Netflix and originating from United States of America,* of the genre [MASK].

Thus, in total 19 prompt variations are considered for each movie.

4. Evaluation

4.1. Dataset

In order to test our approach, we use the BERT [1] and RoBERTa large [20] pre-trained models. The test dataset we are using is a subset of ML25M from IMDB [21]. ML25M contains title and ground truth genre classification of a range of 54,758 movies. A subset of this dataset was then assembled, as those movies for which the Wikidata properties as listed in Table 1 were present in full. This resulted in a test set of 9,596 movie titles. The Wikidata properties, and thus the corresponding data subset, were selected as a compromise between a large dataset, and a diverse set of domain-relevant Wikidata properties, following exploratory analysis of the ML25M dataset.

4.2. Results

Table 2 lists the recall@n scores for each of the prompts described in Section 3.2, for the BERT and RoBERTa large models respectively. For a given model and prompt, recall@1 and recall@5 values for each movie are calculated as the fraction of movie ground-truth genres predicted in the highest ranked *n* PLM mask predictions. The aggregated recall@n values reported in Table 2 are the micro-averaged recall@n scores across all movies in the test dataset, with respect to the model and prompt referenced.

With reference to Table 2, certain variations of the enriched probes showed greater R@n scores that the non-enriched case, for both the BERT and RoBERTa large models, across verbalisation strategies. We compare the statistical significance of the R@n outcomes, of the highest performing enriched prompts (bolded) against the non-enriched case, via a one-tailed, directional, dependent t-test. Where the null hypothesis is that average of the R@n differences is 0, and the alternative hypothesis is that the average of the R@n differences is non-zero, biased towards the selected enriched probe. A significance of 0.05 is applied. With reference to the p-values given in Table 3, we can affirm with statistical significance that the enriched prompts are more performant overall.

Prompt	BERT				RoBERTa large			
	Verbalisation Strategy A		Verbalisation Strategy B		Verbalisation Strategy A		Verbalisation Strategy B	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
non-enriched	0.136	0.448	0.136	0.448	0.065	0.198	0.065	0.198
1	0.139	0.476	0.153	0.487	0.096	0.264	0.076	0.204
2	**0.161**	**0.515**	**0.161**	**0.498**	0.114	0.297	0.016	0.068
3	0.092	0.423	0.065	0.305	0.057	0.180	0.005	0.031
4	0.024	0.226	0.036	0.258	0.010	0.100	0.004	0.038
5	0.017	0.176	0.011	0.090	0.034	0.115	0.012	0.043
6	0.004	0.104	0.020	0.062	0.013	0.053	0.012	0.044
7	0.055	0.320	0.021	0.214	**0.191**	**0.556**	0.203	0.534
8	0.047	0.250	0.006	0.065	0.163	0.466	0.142	0.389
9	0.020	0.140	0.014	0.083	0.184	0.536	**0.210**	**0.576**

Table 2
Recall@n scores for the Bert, RoBERTa large and the movie data subset, averaged over all movies. Verbalisation strategy A and B prompts consist of comma separated and 'and' separated WikiData information, respectively, as described in Section 3.2. The greatest recall@n scores are highlighted in bold.

4.3. Discussion

The results and analysis of Section 4.2 give support to the position that, when considered en-masse, enrichment of prompts with domain-relevant information from Wikidata can improve cloze-style genre prediction in the movie domain. This is the case for both of the investigated verbalisation strategies.

				mean difference	test statistic	p-values
BERT	**Verbalisation Strategy A**	Prompt 2	R@1	0.0245	8.33	0^*
		Prompt 2	R@5	0.0672	21.0	0^*
	Verbalisation Strategy B	Prompt 2	R@1	0.0252	8.47	0^*
		Prompt 2	R@5	0.0506	15.2	0^*
RoBERTa large	**Verbalisation Strategy A**	Prompt 7	R@1	0.125	43.0	0^*
		Prompt 7	R@5	0.358	86.5	0^*
	Verbalisation Strategy B	Prompt 9	R@1	0.144	47.7	0^*
		Prompt 9	R@5	0.378	92.5	0^*

Note: * denotes that the p-value is 0 to at least 3 significant figures.

Table 3
Results for separate dependent one-tailed t-tests under the alternative hypothesis that the average difference between the enriched and non-enriched prompts is non-zero in favour of the enriched case. A p−value less than 0.05 means that we accept the alternative hypothesis with a 5% chance of Type I error.

It is noteworthy, however, that the BERT and RoBERTa large models behave very differently in terms of both their non-enriched performance and their performance when subject to varying levels of enrichment. This is demonstrative of the potential for PLM improvement via prompt enrichment as being highly specific to the model in question. BERT demonstrates optimum recall performance in aggregate for those enriched prompts with relatively low levels of information enrichment, followed by a very rapid reduction in recall@n for further enriched prompts. Whereas RoBERTa large demonstrates fluctuating performance relative to the non-enriched prompt, with the greatest performance shown in the more information-rich prompts.

It is beyond the scope of this paper to disentangle the role of information variety and the specific information types themselves, as to the influence on prediction outcomes. However, there are preliminary indications of complex interactions. For example, as shown in Table 2, prompt 7 (verbalisation strategy A) applied to RoBERTa large shows a huge spike in improved performance over the worst performing prompt 6, which adds the *release date* information. Analysis of a verbalisation strategy A prompt enriched only by *release date* alone, explains a large portion of the improvement (recall@1 = 0.167, recall@5 = 0.48). However, the overall context provided by prompt 7 results in the best performance overall: A one-tail dependent t test between prompt 7 and the case of enrichment by only *release date*, demonstrates significant non-zero differences, in the direction of greater prompt 7 performance for each of recall@1 and recall@5. Both tests reporting a p−value close to 0, with respect to a 0.05 significance. Accordingly, the results are suggestive of further investigative work being required to understand better the interactive effect of information enrichment on whatever model, domain, and task to which such enriched prompts may be applied.

5. Conclusion

Given that PLMs are limited in performance for domain-specific cloze-style question answering prompts, in this paper we examine how adding additional context to naive prompts from KGs can improve the performance of PLMs on a movie genre prediction task. Through our experiments, we show a statistically significant improvement in recall on prompts enriched with information from the Wikidata KG in comparison to non-enriched prompts on the BERT and RoBERTa large PLMs.

As future work, we plan to expand our study to include more domains such as books, music etc. to better understand domain-specific optimum characteristics for enrichment, and cover the same domains as similar previous work [13]. Additionally, we look forward to enriching prompts using web entities [22]. These entities are embedded in HTML pages on the web using Microformat, Microdata and RDFa from the Common Crawl web corpus, the largest and most up-to-date data web corpus available to the public. As more and more websites embed structured data describing for instance products, people, organizations, places, events, resumes, and cooking recipes, the engineered prompts covered domain-specific knowledge that is not present in the encyclopedic Wikidata.

Acknowledgements

We would like to thank the International Semantic Web Summer School 2022, which initiated the collaboration between the authors in producing this paper. This work was funded in-part by: 'Culturally Aware AI' funded by NWO, the ANR-19-CE23-0014 DeKaloG project (CE23 - Intelligence artificielle) and the CominLabs MiKroloG project, Samsung Research UK. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv abs/1810.04805 (2019).

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] L. Asprino, W. Beek, P. Ciancarini, F. v. Harmelen, V. Presutti, Observing lod using equivalent set graphs: it is mostly flat and sparsely linked, in: International Semantic Web Conference, Springer, 2019, pp. 57–74.

[4] P. Zhang, T. Chai, Y. Xu, Adaptive prompt learning-based few-shot sentiment analysis, ArXiv abs/2205.07220 (2022).

[5] T. Kasahara, D. Kawahara, N. Tung, S. Li, K. Shinzato, T. Sato, Building a personalized dialogue system with prompt-tuning, ArXiv abs/2206.05399 (2022).

[6] K. Qi, H. Wan, J. Du, H. Chen, Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 1910–1923.
[7] S. Hu, N. Ding, H. Wang, Z. Liu, J. Wang, J. Li, W. Wu, M. Sun, Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2225–2240.
[8] B. Heinzerling, K. Inui, Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries, ArXiv abs/2008.09036 (2021).
[9] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials, 2019, pp. 15–18.
[10] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, arXiv preprint arXiv:2107.13586 (2021).
[11] T. Mickus, D. Paperno, M. Constant, K. van Deemter, What do you mean, bert? assessing bert as a distributional semantics model, ArXiv abs/1911.05758 (2019).
[12] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, 2019. URL: https://arxiv.org/abs/1911.12543. doi:10.48550/ARXIV.1911.12543.
[13] G. Penha, C. Hauff, What does bert know about books, movies and music? probing bert for conversational recommendation, Fourteenth ACM Conference on Recommender Systems (2020).
[14] T. Shin, Y. Razeghi, R. L. L. IV, E. Wallace, S. Singh, Eliciting knowledge from language models using automatically generated prompts, ArXiv abs/2010.15980 (2020).
[15] S. Min, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Noisy channel language model prompting for few-shot text classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5316–5330.
[16] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 255–269.
[17] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, arXiv preprint arXiv:2104.08786 (2021).
[18] P. West, C. Bhagavatula, J. Hessel, J. D. Hwang, L. Jiang, R. L. Bras, X. Lu, S. Welleck, Y. Choi, Symbolic knowledge distillation: from general language models to commonsense models, ArXiv abs/2110.07178 (2021).
[19] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 3027–3035.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[21] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, Acm transactions on interactive intelligent systems (tiis) 5 (2015) 1–19.
[22] H. Mühleisen, C. Bizer, Web data commons-extracting structured data from two large web corpora, in: LDOW, 2012.