Distilling Hypernymy Relations from Language Models: On the Effectiveness of Zero-Shot Taxonomy Induction

Devansh Jain[♠*] **Luis Espinosa Anke**[◇]
[♠] Department of Computer Science and Information Systems, BITS Pilani, India
[◇] CardiffNLP (Cardiff University) - AMPLYFI
f20180798@pilani.bits-pilani.ac.in
espinosa-ankel@cardiff.ac.uk

Abstract

In this paper, we analyze zero-shot taxonomy learning methods which are based on distilling knowledge from language models via prompting and sentence scoring. We show that, despite their simplicity, these methods outperform some supervised strategies and are competitive with the current state-of-the-art under adequate conditions. We also show that statistical and linguistic properties of prompts dictate downstream performance[1].

1 Introduction

Taxonomy learning (TL) is the task of arranging domain terminologies into hierarchical structures where terms are nodes and edges denote *is-a* (hypernymic) relationships (Hwang et al., 2012). Domain-specific concept generalization is at the core of human cognition (Yu et al., 2015), and a key enabler in NLP tasks where inference and reasoning are important, e.g.: semantic similarity (Pilehvar et al., 2013; Yu and Dredze, 2014), WSD (Agirre et al., 2014) and, more recently, QA (Joshi et al., 2020) and NLI (Chen et al., 2020).

Earlier approaches to taxonomy learning focused on mining lexico-syntactic patterns from candidate (hyponym, hypernym) pairs (Hearst, 1992; Snow et al., 2004; Kozareva and Hovy, 2010; Boella and Di Caro, 2013; Espinosa-Anke et al., 2016), clustering (Yang and Callan, 2009), graph-based methods (Fountain and Lapata, 2012; Velardi et al., 2013) or word embeddings (Fu et al., 2014; Yu et al., 2015). These methods, which largely rely on hand-crafted features, are still relevant today, and complement modern approaches exploiting language models (LMs), either via sequence classification (Chen et al., 2021), or combining contextual, distributed, and lexico-syntactic features (Yu et al., 2020). In parallel, several works have recently focused on using LMs as zero-shot tools for solving NLP tasks, e.g., commonsense, relational and analogical reasoning (Petroni et al., 2019; Bouraoui et al., 2020; Ushio et al., 2021; Paranjape et al., 2021), multiword expression (MWE) identification (Espinosa-Anke et al., 2021; Garcia et al., 2021), QA (Shwartz et al., 2020; Banerjee and Baral, 2020), domain labeling (Sainz and Rigau, 2021), or lexical substitution and simplification (Zhou et al., 2019). Moreover, by tuning and manipulating natural language queries (often referred to as *prompts*), impressive results have been recently obtained on tasks such as semantic textual similarity, entailment, or relation classification (Shin et al., 2020; Qin and Eisner, 2021).

In this paper, we evaluate LMs on TL benchmarks using prompt-based and sentence-scoring techniques, and find not only that they are competitive with common approaches proposed in the literature (which are typically supervised and/or reliant on external resources), but that they achieve state-of-the-art results in certain domains.

2 Methodology

We follow Ushio et al. (2021) and define a prompt generation function $\tau_p(t_1, t_2)$ which maps a pair of terms and a prompt type p to a single sentence. For instance,

$$\tau_{kind}(\text{“physics”}, \text{“science”}) = \text{“physics is a kind of science”}$$

Then, given a terminology $\mathcal{T}$, the goal is to, given an input term $t \in \mathcal{T}$, retrieve its top k most likely hypernyms, (in our experiments, $k \in \{1, 3, 5\}$), using either masked language model (MLM) prompting (§2.1), or sentence-scoring (§2.2).

2.1 MLM Prompting

RestrictMLM Petroni et al. (2019) introduced a “fill-in-the-blanks” approach based on cloze state-

[*] Work done during an internship at CardiffNLP.
[1] Code available at https://github.com/devanshrj/zero-shot-taxonomy.

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 151 - 156
July 14-15, 2022 ©2022 Association for Computational Linguistics

ments (or *prompts*) to extract relational knowledge from pretrained LMs. The intuition being that an LM can be considered to "know" a fact (in the form of a <*subject, relation, object*> triple) such as <*Madrid, capital-of, Spain*> if it can successfully predict the correct words when queried with prompts such as "Madrid is the capital of [MASK]". We extend this formulation to define a hypernym retrieval function $f_R(\cdot)$ as follows:

$$f_R(p, t, \mathbf{T}) = P(\texttt{[MASK]}|\tau_p(t, \texttt{[MASK]})) * \mathbf{T} \quad (1)$$

where p is a prompt type, and $\mathbf{T}$ is a one-hot encoding of the terms $\mathcal{T}$ in the LM's vocabulary. We follow previous works (Petroni et al., 2019; Kassner et al., 2021) and restrict the output probability distribution since this task requires the construction of a lexical taxonomy starting from a fixed vocabulary.

PromptMLM For completeness, we also report results for an unrestricted variant of *RestrictMLM*, where the LM's entire vocabulary is considered.

2.2 LMScorer

Factual (and true) information such as "Trout is a type of fish" should be scored higher by a LM than fictitious information such as "Trout is a type of mammal". The method for scoring a sentence depends on the type of LM used.

Causal Language Models Given a sentence $\mathbf{W}$, causal LMs ($\mathcal{C}$) predict token w_i using only past tokens $\mathbf{W}_{<i}$. Thus, a likelihood score can be estimated for each token w_i from the LM's next token prediction. The corresponding scores are then aggregated to yield a score for sentence $\mathbf{W}$.

$$s_{\mathcal{C}}(\mathbf{W}) = \exp\left(\sum_{i=1}^{|\mathbf{W}|} logP_{\mathcal{C}}(w_i|\mathbf{W}_{<i})\right) \quad (2)$$

Masked Language Models Given a sentence $\mathbf{W}$, masked LMs ($\mathcal{M}$) replace w_i by [MASK] and predict it using past and future tokens. Thus, a pseudo-likelihood score can be computed for each token w_i by iteratively masking it and using the LM's masked token prediction (Wang and Cho, 2019; Salazar et al., 2020). The corresponding scores are then aggregated to yield a score for sentence $\mathbf{W}$.

$$s_{\mathcal{M}}(\mathbf{W}) = \exp\left(\sum_{i=1}^{|\mathbf{W}|} logP_{\mathcal{M}}(w_i|\mathbf{W}_{\setminus i})\right) \quad (3)$$

Given the above, we can cast TL as a sentence-scoring problem by evaluating the natural fluency of hypernymy-eliciting sentences. Specifically, for each term t, we score the sentences generated using $\tau_p(\cdot)$ with every other term t' in the terminology. We then select the term-pair with the highest sentence score and assume that the corresponding term t' is a hypernym of t. Formally, we define a hypernym selection function $f_S(\cdot)$ as follows:

$$f_S(p, t, \mathcal{T}) = \arg\max_{t' \in \mathcal{T} \setminus t}[s(\tau_p(t, t'))] \quad (4)$$

where s refers to the scoring function determined by the LM used.

3 Experimental setup

This section covers the datasets and prompts we use in our experiments[2], as well as the different LMs we consider. Concerning evaluation metrics, we report standard precision (P), recall (R) and F-score at the *edge level* (Bordea et al., 2016).

Dataset Details We evaluate our proposed approaches on datasets belonging to two TL SemEval tasks (*TExEval-1*, Bordea et al. (2015) and *TExEval-2*, Bordea et al. (2016)). Following recent literature, we consider the *equipment* taxonomy from *TExEval-1* and the English-language *environment*, *science* and *food* taxonomies from *TExEval-2*. For the *science* taxonomy, our results are based on an *average of the 3 subsets*, which is in line with previous work. Since these datasets do not come with training data, they are well suited for unsupervised approaches.

Domain	Source	V	E
environment	Eurovoc	261	261
science	Combined	453	465
	Eurovoc	125	124
	WordNet	429	452
food	Combined	1556	1587
equipment	Combined	612	615

Table 1: Taxonomies statistics. Vertices (V) and Edges (E) are often used as structural measures.

[2]We use PyTorch and the `transformers` library (Wolf et al., 2020), as well as `mlm-scoring` (Salazar et al., 2020) (`https://github.com/awslabs/mlm-scoring`).

Prompts We use the following prompts:

- *gen.*: $[t_2]$ is more general than $[t_1]$.
- *spec.*: $[t_1]$ is more specific than $[t_2]$.
- *type*: $[t_1]$ is a type of $[t_2]$.

$gen.$ and $spec.$ prompts are hand-crafted templates to encode, in a general way, the hypernymy relationship. The choice of the $type$ prompt, however, comes from a set of experiments involving all *LPAQA* (Jiang et al., 2020) prompts under the "*is a subclass of*" category. We do not consider automatic prompt generation techniques (Shin et al., 2020) due to the absence of training data. Note that for each prompt, we replace t_1 with the input term so that the task is always to predict its hypernym.

Language Models We interrogate BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) among masked LMs, and GPT2 (Radford et al., 2019) among causal LMs. For each LM, we consider two variants corresponding to approximately 117M parameters and 345M parameters.

4 Results

Table 2 shows the results on *TExEval-2*'s *science* and *environment*. We compare with the current state of the art (*Graph2Taxo*) (Shang et al., 2020), as well as with other strong baselines such as *TaxoRL* (Mao et al., 2018) and *TAXI* (Panchenko et al., 2016), the highest ranked system in *TExEval-2*. We also compare with *CTP* (Chen et al., 2021) to illustrate the advantages of zero-shot methods vs finetuning. For the *environment* domain, we find that *RestrictMLM* performs similar to *CTP* and *LMScorer* outperforms it. Moreover, all 3 proposed approaches fail to outperform the other baselines. However, in *science*, all 3 of our approaches outperform *CTP*, while our best model (*RestrictMLM*) outperforms *TAXI* and is competitive with *TaxoRL* (ours has higher precison, but lower recall). Note that compared to our zero-shot approaches, these methods are either supervised, expensive to train or take advantage of external taxonomical resources such as WordNet, or lexico-syntactic patterns mined from the web using different hand-crafted heuristics.

We also show results for *TExEval-1*'s *equipment* and *TExEval-2*'s *food* (Table 3). Both datasets are considerably larger than *environment* and *science*. We compare with the corresponding highest ranked system, namely *TAXI* for *food*, and *INRIASAC* (Grefenstette, 2015) for *equipment*. For both domains, all 3 of our approaches outperform the corresponding *TExEval* best-performing systems. This suggests that zero-shot TL with LMs is robust, easily scalable and feasible on large taxonomies. However, a clear bottleneck for prompt-based methods is that only single-token terms can be predicted (using a single `[MASK]` token), making this approach a lower bound for TL.

	environment			*science*		
Model	P	R	F	P	R	F
TAXI	33.8	26.8	29.9	35.2	35.3	35.2
TaxoRL	32.3	32.3	32.3	37.9	37.9	37.9
Graph2Taxo	89.0	24.0	**37.0**	84.0	30.0	**44.0**
CTP	23.1	23.0	23.0	29.4	28.8	29.1
PromptMLM	19.2	19.2	19.2	34.4	32.0	33.1
RestrictMLM	23.0	23.0	23.0	39.3	36.7	37.9
LMScorer	26.4	26.4	26.4	33.1	30.7	31.8

Table 2: Comparison of our best performing methods with previous work (*environment* and *science*).

	food			*equipment*		
Model	P	R	F	P	R	F
TExEval	13.2	25.1	17.3	51.8	18.8	27.6
PromptMLM	23.2	22.6	22.9	29.4	29.3	29.4
RestrictMLM	25.2	24.6	**24.9**	38.4	38.2	**38.3**
LMScorer	25.2	24.6	**24.9**	37.7	37.6	37.6

Table 3: Comparison of our best configurations with the best TExEval systems on *food* and *equipment*.

5 Analysis

In this section, we provide an in-depth analysis of our approaches, including comparison of LMs and statistical and semantic properties of prompts.

LM Comparison Table 4 compares the best configuration for each LM. We can immediately see that a conservative approach (i.e., $k = 1$ with the *type* prompt) almost always yields the best F-score. Another important conclusion is that, among MLMs, BERT-Large performs best across the board, with BERT generally outperforming RoBERTa, a finding in line with previous works (Shin et al., 2020). Concerning causal LMs, GPT-2 Medium outperforms its smaller counterpart as well as both MLMs for sentence-scoring.

Sensitivity to Prompts There is interest in understanding models' sensitivity to prompts and

		environment				*science*				*food*				*equipment*			
Method	LM	(p,k)	P	R	F	(p,k)	P	R	F	(p,k)	P	R	F	(p,k)	P	R	F
PromptMLM	BERT-Base	$(t,1)$	18.8	18.8	18.8	$(t,1)$	30.2	28.1	29.1	$(t,1)$	20.9	20.4	20.6	$(t,1)$	29.4	29.3	29.4
	BERT-Large	$(t,1)$	19.2	19.2	19.2	$(t,1)$	34.4	32.0	33.1	$(t,1)$	23.2	22.6	22.9	$(t,1)$	28.4	28.3	28.4
	RoBERTa-Base	$(t,1)$	18.0	18.0	18.0	$(t,1)$	24.5	23.0	23.7	$(t,1)$	18.5	18.0	18.2	$(t,1)$	26.3	26.2	26.3
	RoBERTa-Large	$(t,1)$	18.0	18.0	18.0	$(t,1)$	28.1	26.2	27.1	$(t,1)$	20.3	19.8	20.0	$(t,1)$	28.4	28.3	28.4
RestrictMLM	BERT-Base	$(t,1)$	23.0	23.0	23.0	$(t,1)$	35.8	33.5	34.6	$(t,1)$	22.8	22.2	22.5	$(t,1)$	38.4	38.2	38.3
	BERT-Large	$(t,1)$	21.8	21.8	21.8	$(t,1)$	39.3	36.7	**37.9**	$(t,1)$	25.2	24.6	**24.9**	$(t,1)$	37.9	37.7	**37.8**
	RoBERTa-Base	$(t,1)$	5.4	5.4	5.4	$(t,1)$	11.0	10.6	10.8	$(t,1)$	9.3	9.1	9.2	$(t,1)$	0.0	0.0	0.0
	RoBERTa-Large	$(t,1)$	8.4	8.4	8.4	$(t,1)$	12.3	11.8	12.0	$(t,1)$	10.7	10.5	10.6	$(t,1)$	0.0	0.0	0.0
LMScorer	BERT-Base	$(t,1)$	20.3	20.3	20.3	$(t,1)$	15.2	14.4	14.8	$(t,3)$	6.8	19.7	10.1	$(t,3)$	7.5	22.4	11.2
	BERT-Large	$(t,3)$	13.7	41.0	20.5	$(t,1)$	13.0	12.4	12.6	$(t,1)$	13.9	13.6	13.7	$(t,1)$	15.2	15.1	15.1
	RoBERTa-Base	$(g,3)$	7.7	23.0	11.5	$(t,3)$	5.5	15.7	8.1	$(t,3)$	2.5	7.2	3.7	$(t,5)$	4.2	21.0	7.0
	RoBERTa-Large	$(t,3)$	11.1	33.3	16.7	$(t,1)$	13.6	12.8	13.2	$(t,3)$	3.6	10.6	5.4	$(t,3)$	9.2	27.5	13.8
	GPT-2 Base	$(t,1)$	24.9	24.9	24.9	$(t,1)$	29.3	27.4	28.3	$(t,1)$	21.0	20.5	20.7	$(t,1)$	36.8	36.6	36.7
	GPT-2 Medium	$(t,1)$	26.4	26.4	**26.4**	$(t,1)$	33.1	30.7	31.8	$(t,1)$	25.2	24.6	**24.9**	$(t,1)$	37.7	37.6	37.7

Table 4: Comparison of best configuration for each LM and proposed approach. (p,k) refers to the prompt and top-k combination that gives the best results for that setting, where $p = g$ for *gen.*, s for *spec.* and t for *type* prompt.

whether frequency can explain downstream performance in lexical semantics tasks (Chiang et al., 2020). In the context of prompt vs. performance correlation, we find that prompt-based downstream performance on TL can be attributed to: (1) syntactic completeness and (2) semantic correctness. For (1), we find that prompts that are syntactically more complete (e.g., "*[X] is a type of [Y]*" vs "*[X] is a type [Y]*", the difference being the prepositional phrase) perform better. For (2), we find that prompts that unambiguously encode hypernymy are also better (i.e., the *type* prompt, as opposed to other noise-inducing templates such as "*is a*" or "*is kind of*"). Finally, out of the cleanest prompts, the most frequent in pretraining corpora are the most competitive. Table 5 confirms the intuition that the *type* prompt is not only unambiguous, but also highly frequent when compared to similar (noise-free and syntactically complete) prompts.

Prompt	*avg* F	Frequency
is a type of	25.5	14,503
is the type of	24.2	809
is a kind of	23.6	2,934
is a form of	22.1	9,518
is one form of	17.9	124
is a	7.4	9,328,426
is a type	1.0	15,085

Table 5: Domain-wise average F-score of LPAQA prompts and their frequency in BERT's pretraining corpora.

Single-Token vs Multi-Token Hypernyms Table 6 compares *F-score* on original terminology vs filtered terminology, where filtered terminology contains only the terms that have single-token hypernyms. The results show that % Increase in *F-score* is inversely proportional to the % Retained. This can be explained by the fact that smaller % of terms retained implies higher % of multi-token hypernyms in the original dataset that cannot be predicted using prompting. Thus, the increase in *F-score* by removing such hypernyms should increase as the % Retained decreases.

Domain	Total Terms	% Retained	% Increase
environment	261	29.89	2.32
equipment	612	44.77	1.24
science	452	53.32	0.90
science_ev	125	52.80	0.89
food	1555	59.55	0.57
science_wn	370	69.73	0.51

Table 6: Comparison of *F-score* on original terminology vs filtered terminology. % Retained refers to the percentage of terms that have single-token hypernyms and are thus retained for the filtered dataset. % Increase shows the increase in *F-score* on filtered dataset compared to *F-score* on original dataset.

6 Conclusion and Future Work

We have presented a study of different LMs under different settings for zero-shot taxonomy learning. Compared with computationally expensive and highly heuristic methods, our zero-shot alternatives prove remarkably competitive. For the future, we could explore multilingual signals and the integration of traditional word embeddings with contextual representations.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–162.

Guido Boella and Luigi Di Caro. 2013. Supervised learning of syntactic contexts for uncovering definitions and extracting hypernym relations in text databases. In *Machine learning and knowledge discovery in databases*, pages 64–79. Springer.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*. 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.

Catherine Chen, Kevin Lin, and D. Klein. 2021. Constructing taxonomies from pretrained language models. In *NAACL*.

Mingda Chen, Zewei Chu, Karl Stratos, and Kevin Gimpel. 2020. Mining knowledge for natural language inference from wikipedia categories. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3500–3511.

Hsiao-Yu Chiang, Jose Camacho-Collados, and Zachary Pardos. 2020. Understanding the source of semantic regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 119–131.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Luis Espinosa-Anke, Joan Codina-Filbá, and Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1406–1417.

Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016. Extasem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of AAAI*, Phoenix, USA.

Trevor Fountain and Mirella Lapata. 2012. Taxonomy induction using hierarchical random graphs. In *Proceedings of NAACL*, pages 466–476. Association for Computational Linguistics.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*, volume 1.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564.

Gregory Grefenstette. 2015. Inriasac: Simple hypernym extraction methods. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 911–914.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545.

Sung Ju Hwang, Kristen Grauman, and Fei Sha. 2012. Semantic kernel forests from multiple taxonomies. In *Advances in Neural Information Processing Systems*, pages 1718–1726.

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

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *arXiv preprint arXiv:2004.12006*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of EMNLP*, pages 1110–1118.

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

Yuning Mao, Xiang Ren, J. Shen, X. Gu, and Jiawei Han. 2018. End-to-end reinforcement learning for automatic taxonomy induction. In *ACL*.

Alexander Panchenko, Stefano Faralli, E. Ruppert, Steffen Remus, Hubert Naets, Cedric Fairon, Simone Paolo Ponzetto, and Chris Biemann. 2016. Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *SemEval@NAACL-HLT*.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Prompting contrastive explanations for commonsense reasoning tasks. *arXiv preprint arXiv:2106.06823*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, A. Bakhtin, Yuxiang Wu, Alexander H. Miller, and S. Riedel. 2019. Language models as knowledge bases? In *EMNLP*.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351, Sofia, Bulgaria.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Oscar Sainz and German Rigau. 2021. Ask2transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*, pages 44–52.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *ACL*.

Chao Shang, Sarthak Dash, Md. Faisal Mahbub Chowdhury, Nandana Mihindukulasooriya, and A. Gliozzo. 2020. Taxonomy construction of unseen domains via graph-based cross-domain knowledge transfer. In *ACL*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

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.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Asahi Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies? In *Proceedings of the ACL-IJCNLP 2021 Main Conference*. Association for Computational Linguistics.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn Reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hui Yang and Jamie Callan. 2009. A metric-based framework for automatic taxonomy induction. In *Proceedings of ACL/IJCNLP*, pages 271–279. Association for Computational Linguistics.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *ACL (2)*, pages 545–550.

Yue Yu, Yinghao Li, Jiaming Shen, Haoyang Feng, Jimeng Sun, and Chao Zhang. 2020. Steam: Self-supervised taxonomy expansion with mini-paths. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of IJCAI*, pages 1390–1397.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.