of next word $P(x_t|[\boldsymbol{x}_{<t}; \boldsymbol{x}_{\text{diagnosis}}]; \theta)$ by appending self-diagnosis textual input to the original input as mentioned above. These two probability distributions for the next token can be combined to suppress the undesired attribute.

Dataset Construction Schick and Schütze (2021) propose to use pre-trained LMs to generate datasets given certain instructions. As an example, suppose we have an unlabeled dataset in which each sample is a sentence. If we want to construct a dataset containing pairs of semantically similar sentences, then we can use the following template for each input sentence: "Write two sentences that mean the same thing. `[X] [Z]`" and attempt to generate a sentence that shares the same meaning as the input sentence.

8.10 Resources

We also collect some useful resources for different prompt-based applications.

Dataset Some datasets specifically designed for few-shot and zero-shot learning are shown in Tab. 9.

Task	Dataset	Setting	URL
Commonsense Reasoning	Pronoun Disambiguation Problems [93]	Zero	https://cs.nyu.edu/ davise/papers/...
	Winograd Schema Challenge [93]	Zero	https://cs.nyu.edu/ davise/papers/...
	CPRAG-102 [39]	Zero	https://github.com/aetting/lm-diagnostics
Linguistic Capacity Probing	WNLaMPro [150]	Zero	https://github.com/timoschick/...
	ROLE-88 [39]	Zero	https://github.com/aetting/lm-diagnostics
	NEG-136 [39]	Zero	https://github.com/aetting/lm-diagnostics
Fact Probing	LAMA [133]	Zero	https://dl.fbaipublicfiles.com/LAMA/...
	Negated LAMA [74]	Zero	https://github.com/norakassner/LAMA...
	Misprimed LAMA [74]	Zero	https://github.com/norakassner/LAMA...
	X-FACTR [66]	Zero	https://x-factr.github.io/
	LAMA-TREx-easy-hard [203]	Zero	https://github.com/princeton-nlp/...
Text Classification	FLEX [15]	Zero,Few	https://github.com/allenai/flex
	FewGLUE [154]	Few	https://github.com/timoschick/fewglue
General Conditional Gen.	REALTOXICITYPROMPTS [47]	Zero	https://allenai.org/data/...
	Natural-Instructions [120]	Few,Full	https://instructions.apps.allenai.org/

Table 9: Few-shot and zero-shot datasets for prompt-based learning.

Prompts As shown in Tab. 10, we collect existing commonly-used prompts designed manually, which can be regarded as off-the-shelf resource for future research and applications.

9 Prompt-relevant Topics

What is the essence of prompt-based learning and how does it relate to other learning methods? In this section, we connect prompt learning with other similar learning methods.

Ensemble Learning *Ensemble learning* (Ting and Witten, 1997; Zhou et al., 2002) is a technique that aims to improve the performance of a task by taking advantage of the complementarity of multiple systems. Generally, the different systems used in an ensemble result from different choices of architectures, training strategies, data ordering, and/or random initialization. In prompt ensembling (§6.1), the choice of prompt templates becomes another way to generate multiple results to be combined. This has the clear advantage that this does not necessarily require training the model multiple times. For example, when using discrete prompts, these prompts can simply be changed during the inference stage (Jiang et al., 2020c).

Few-shot Learning *Few-shot learning* aims to learn a machine learning system in the data-scarce scenarios with few training samples. There are a wide variety of methods to achieve few-shot learning including model agnostic meta-learning (Finn et al., 2017b) (learning features rapidly adaptable to new tasks), embedding learning (Bertinetto et al., 2016) (embedding each sample in a lower-dimensional space where similar samples are close together), memory-based learning (Kaiser et al., 2017) (representing each sample by a weighted average of contents from the memory) etc. (Wang et al., 2020). Prompt augmentation can be regarded as another way to achieve few-shot learning (a.k.a. priming-based few-shot learning (Kumar and Talukdar, 2021)). Compared to previous methods, prompt augmentation directly prepends several labeled samples to the currently-processed sample elicit knowledge from pre-trained LMs even without any parameter tuning.

Task	Example Prompt-Answer	Resource
Fact Probing	**Prompt** Adolphe Adam died in `[Z]`. **Answer** $\mathcal{V}$ **Prompt** iPod Touch is produced by `[Z]`. **Answer** $\mathcal{V}$ **Prompt** The official language of Mauritius is `[Z]`. **Answer** $\mathcal{V}$	LAMA dataset LPAQA dataset X-FACTR dataset
Text Classificatin	**Prompt** Which of these choices best describes the following document? "`[Class A]`", "`[Class B]`", "`[Class C]`". `[X][Z]` **Answer** `[Class A]`, `[Class B]`, `[Class C]` **Prompt** How is the text best described? : "`[Class A]`", "`[Class B]`" , or "`[Class C]`". `[X][Z]` **Answer** `[Class A]`, `[Class B]`, `[Class C]` **Prompt** This passage is about `[Z]`: `[X]` **Answer** `[Class A]`, `[Class B]`, `[Class C]` **Prompt** `[X]`. Is this review positive? `[Z]` **Answer** Yes, No **Prompt** `[X]` It was `[Z]`. **Answer** great, terrible	Meta [202]
Natural Language Inference	**Prompt** `[X1]`? `[Z]`, `[X2]` **Answer** Yes, No, Maybe **Prompt** `[X1]` `[Z]`, `[X2]` **Answer** Yes, No, Maybe	
Commonsense Reasoning	**Prompt** The trophy doesn't fit into the brown suitcase because `[Z]` is too large. **Answer** trophy, suitcase **Prompt** Ann asked Mary what time the library closes, because `[Z]` had forgotten. **Answer** Ann, Mary	PDP dataset WSC dataset CPRAG-102 dataset
Linguistic Knowledge Probing	**Prompt** A robin is a `[Z]`. **Answer** bird, tree **Prompt** A robin is not a `[Z]`. **Answer** bird, tree **Prompt** New is the opposite of `[Z]`. **Answer** old, young, current	WNLaMPro dataset ROLE-88 dataset NEG-136 dataset
Named Entity Recognition	**Prompt-Pos** `[X]` `[Span]` is a `[Z]` entity. **Prompt-Neg** `[X]` `[Span]` is not a named entity. **Answer** person, location, organization, miscellaneous **Prompt-Pos** The entity type of `Span` is `[Z]`. **Prompt-Neg** `[X]` The entity type of `[Span]` is none entity. Answer person, location, organization, miscellaneous	TemplateNER [29]
Question Answering	**Prompt** `[Question]` `[Passage]` `[Z]` **Prompt** `[Passage]` According to the passage, `[Question]` `[Z]` **Prompt** Based on the following passage, `[Question]` `[Z]`. `[Passage]`	
Summarization	**Prompt** Text: `[X]` Summary: `[Z]` **Prompt** `[X]` TL;DR: `[Z]` **Prompt** `[X]` In summary, `[Z]`	BARTScore [193]
Machine Translation	**Prompt** French: `[French sentence]` English: **Prompt** A French sentence is provided: `[French sentence]` The French translator translates the sentence into English: `[Z]` **Prompt** `[French sentence]` = `[Z]`	

Table 10: Commonly used prompts and answers for different tasks. `[X]` and `[Z]` denote slots for input and answer respectively. $\mathcal{V}$ denotes the vocabulary of the LM. More prompts for each task can be found using the **Resource** column.

Prompt Concept	Relevant Topic	Commonality	Peculiarity
Prompt Ensembling [68; 153]	Ensemble Learning [171; 204]	Combine results of multiple systems to get better performance	In prompt ensembling, multiple predictions result from different prompt variants. This contrasts with architecture or feature variations, each of which requires separate training.
Prompt Augmentation [16; 46]	Few-shot Learning [160; 42]	Use few examples to learn generalized rules	Prompt augmentation is a specific subset of few-shot learning.
	Larger-context Learning [18; 53]	Introduce larger context to aid the learning process	Additional information introduced in larger-context learning is not necessarily the labeled data.
Discrete Prompt Search [68; 159]	Query reformulation [123; 123]	Reformulate the input into a query form	Query reformulation commonly focuses on information extraction and question answering tasks, while prompt learning can be applied to a variety of NLP tasks
Discrete Prompt Fine-tuning [46]	QA-based multi-task learning [115; 97]	Reformulate many tasks into an QA form	QA-based formulations aim to solve different tasks through question answering, while prompting additionally targets full use of pre-trained models.
Continuous Prompt Fine-tuning [103; 36]	Controlled Text Generation [191; 77; 156]	Input is augmented with additional inputs to control the generation process	Controlled generation targets generation of a particular type of text while prompt learning uses prompts to specify the task itself.
Prompt-based downstream task learning [153; 193]	Supervised Attention [101; 165]	Require external hint to remind the model of which part information should be focused on	Research works on supervised attention usually target at salient information from an image or text, while prompt learning aims to utilize relevant knowledge from the pre-trained model.
	Data augmentation [40; 144]	Improving downstream tasks' performance by introducing additional samples	Data augmentation introduce additional training samples in an explicit way while prompts can be regarded as highly-condensed training samples [88].

Table 11: Other research topics relevant to prompting methods.

Larger-context Learning *Larger-context learning* aims to improve the system's performance by augmenting the input with additional contextual information, e.g. retrieved from the training set (Cao et al., 2018) or external data sources (Guu et al., 2020). Prompt augmentation can be regarded as adding relevant labeled samples into the input, but a minor difference is in larger-context learning, the introduced context is not necessarily labeled data.

Query Reformulation *Query reformulation* (Mathieu and Sabatier, 1986; Daumé III and Brill, 2004) is commonly used in information retrieval (Nogueira and Cho, 2017) and question answering tasks (Buck et al., 2017; Vakulenko et al., 2020), which aim to elicit more relevant texts (documents or answers) by expanding the input query with related query terms (Hassan, 2013) or generating paraphrases. There are several commonalities between prompt-based learning and query reformulation, for example (1) both aim to make better use of some existing knowledge bases by asking a right questions (2) the knowledge bases are usually a black-box, not available to the users, so researchers must learn how to probe it optimally based on solely questions.

There are also differences: the knowledge base in traditional query reformulation problems is usually a search engine (Nogueira and Cho, 2017), or QA system (Buck et al., 2017). By contrast, for prompt-based learning, we usually define this knowledge base as an LM, and need to find the appropriate query to elicit an appropriate answer from it. The input reformulation in prompt learning has changed the form of tasks. For example, an original text classification task has been converted into a cloze question problem, therefore bringing additional complexity regarding how to (1) make an appropriate task formulation, and (2) change the modeling framework accordingly. These steps are not required in traditional query formulation. Despite these discrepancies, some methodologies from query reformulation research still can be borrowed for prompt learning, such as decomposing input query into multiple sub-queries (Nogueira et al., 2019), similar to prompt decomposition.

QA-based Task Formulation *QA-based task formulation* aims to conceptualize different NLP tasks as a question-answering problem. (Kumar et al., 2016; McCann et al., 2018) are earlier works that attempt to unify multiple NLP tasks into a QA framework. Later, this idea has been further explored in information extraction (Li et al., 2020; Wu

et al., 2020) and text classification (Chai et al., 2020). These methods are very similar to the prompting methods introduced here in that they use textual questions to specify which task is to be performed. However, one of the key points of prompting methods is how to better use the knowledge in pre-trained LMs, and these were not covered extensively on previous works advocating for QA formulations.

Controlled Generation *Controlled generation* aims to incorporate various types of guidance beyond the input text into the generation model (Yu et al., 2020). Specifically, the guidance signals could be *style tokens* (Sennrich et al., 2016b; Fan et al., 2018), *length specifications* (Kikuchi et al., 2016), *domain tags* (Chu et al., 2017), or any variety of other pieces of information used to control of the generated text. It could also be *keywords* (Saito et al., 2020), *relation triples* (Zhu et al., 2020) or even *highlighted phrases or sentences* (Grangier and Auli, 2018; Liu et al., 2021c) to plan the content of generated texts. In a way, many of the prompting methods described here are a type of controllable generation, where the prompt is usually used to specify the *task itself*. Thus, it is relatively easy to find commonalities between the two genres: (1) both add extra information to the input text for better generation, and these additional signals are (often) learnable parameters. (2) If "controlled generation" is equipped with seq2seq-based pre-trained models (e.g., BART), then it is can be regarded as prompt learning with input-dependent prompts and the *prompt+LM fine-tuning* strategy (§7.2.5), e.g. *GSum* (Dou et al., 2021), where both the prompt's and pre-trained LM's parameters can be tuned.

Also, some clear discrepancies between controlled generation and prompt-based text generation are: (1) In controlled generation work, the control is generally performed over the style or content of the generations (Fan et al., 2018; Dou et al., 2021) while the underlying task remains the same. They don't necessarily require a pre-trained model. In contrast, the main motivation for using prompts for text generation is to specify the task itself and better utilize the pre-trained model. (2) Moreover, most of the current work on prompt learning in text generation shares a dataset- or task-level prompt (Li and Liang, 2021). Only very few works have explored input-dependent ones (Tsimpoukelli et al., 2021). However, this is a common setting and effective in the controlled text generation, which may provide valuable direction for the future work on prompt learning.

Supervised Attention Knowing to pay attention to the important information is a key step when extracting useful information from objects such as long text sequences (Liu et al., 2016; Sood et al., 2020), images (Sugano and Bulling, 2016; Zhang et al., 2020b), or knowledge bases (Yu et al., 2020; Dou et al., 2021)). *Supervised attention* (Liu et al., 2017b) aims to provide explicit supervision over the attention of models based on the fact that completely data-driven attention can overfit to some artifacts (Liu et al., 2017a). In this respect, prompt learning and supervised attention share ideas that both aim to extract salient information with some clues, which need to be provided separately. To solve this problem, supervised attention methods tried to use additional loss functions to learn to predict gold attention on a manually labeled corpus (Jiang et al., 2015; Qiao et al., 2018; Gan et al., 2017). Research on prompt learning may also borrow ideas from this literature.

Data Augmentation Data augmentation is a technique that targets increasing the amount of data that can be used for training by making modifications to existing data (Fadaee et al., 2017; Ratner et al., 2017). As recently observed by (Scao and Rush, 2021), adding prompts can achieve a similar accuracy improvement to the addition of 100s of data points on average across classification tasks, which suggests that using prompts for a downstream task is similar to conducting data augmentation implicitly.

10 Challenges

Although prompt-based learning has shown significant potential among different tasks and scenarios, several challenges remain, some of which we detail below.

10.1 Prompt Design

Tasks beyond Classification and Generation Most existing works about prompt-based learning revolve around either text classification or generation-based tasks. Applications to information extraction and text analysis tasks have been discussed less, largely because the design of prompts is less straightforward. We expect that applying prompting methods to these tasks in the future it will require either reformulating these tasks so that they can be solved using classification or text generation-based methods, or performing effective answer engineering that expresses structured outputs in an appropriate textual format.

Prompting with Structured Information In many NLP tasks, the inputs are imbued with some variety of structure, such as tree, graph, table, or relational structures. How to best express these structures in prompt or answer engineering is a major challenge. Existing works (Chen et al., 2021b) make a step by making prompts with additional marks to encode lexical information, such as entity markings. Aghajanyan et al. (2021) present structured prompts based on hyper text markup language for more fine-grained web text generation. However, moving beyond this to more complicated varieties of structure is largely unexplored, and a potentially interesting research area.

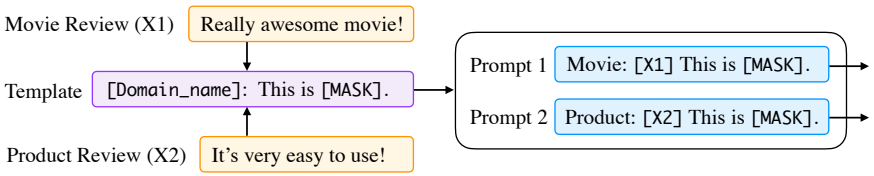


Figure 5: Multi-prompt learning for multi-task, multi-domain or multi-lingual learning. We use different colors to differentiate different components as follows. " " for input text, " " for template, " " for prompt.

Entanglement of Template and Answer The performance of a model will depend on *both* the templates being used and the answer being considered. How to simultaneously search or learn for the best combination of template and answer remains a challenging question. Current works typically select answers before select template (Gao et al., 2021; Shin et al., 2020), but Hambardzumyan et al. (2021) have demonstrated the initial potential of simultaneously learning both.

10.2 Answer Engineering

Many-class and Long-answer Classification Tasks For classification-based tasks, there are two main challenges for answer engineering: (a) When there are too many classes, how to select an appropriate answer space becomes a difficult combinatorial optimization problem. (b) When using multi-token answers, how to best decode multiple tokens using LMs remains unknown, although some multi-token decoding methods have been proposed (Jiang et al., 2020a).

Multiple Answers for Generation Tasks For text generation tasks, qualified answers can be semantically equivalent but syntactically diverse. So far, almost all works use prompt learning for text generation relying solely on a single answer, with only a few exceptions (Jiang et al., 2020c). How to better guide the learning process with multiple references remains a largely open research problem.

10.3 Selection of Tuning Strategy

As discussed in §7, there are a fairly wide variety of methods for tuning parameters of prompts, LMs, or both. However, given the nascent stage of this research field, we still lack a systematic understanding of the tradeoffs between these methods. The field could benefit from systematic explorations such as those performed in the pre-train and fine-tune paradigm regarding the tradeoffs between these different strategies (Peters et al., 2019).

10.4 Multiple Prompt Learning

Prompt Ensembling In prompt ensembling methods, the space and time complexity increase as we consider more prompts. How to distill the knowledge from different prompts remains underexplored. Schick and Schütze (2020, 2021a,b) use an ensemble model to annotate a large dataset to distill the knowledge from multiple prompts.

In addition, how to select ensemble-worthy prompts is also under-explored. For text generation tasks, the study of prompt ensemble learning has not been performed so far, probably because ensemble learning in text generation itself is relatively complicated. To remedy this problem, some recently proposed neural ensembling methods such as *Refactor* (Liu et al., 2021c) could be considered as a method for prompt ensembling in text generation tasks.

Prompt Composition and Decomposition Both prompt composition and decomposition aim to break down the difficulty of a complicated task input by introducing multiple sub-prompts. In practice, how to make a good choice between them is a crucial step. Empirically, for those token (Ma and Hovy, 2016) or span (Fu et al., 2021) prediction tasks (e.g., NER), prompt decomposition can be considered, while for those span relation prediction (Lee et al., 2017) tasks (e.g., entity coreference), prompts composition would be a better choice. In the future, the general idea of de-/composing can be explored in more scenarios.

Prompt Augmentation Existing prompt augmentation methods are limited by the input length, i.e., feeding too many demonstrations to input is infeasible. Therefore, how to select informative demonstrations, and order them in an appropriate is an interesting but challenging problem (Kumar and Talukdar, 2021).

Prompt Sharing All the above considerations refer to the application of prompt in a single task, domain or language. We may also consider *prompt sharing*, where prompt learning is applied to multiple tasks, domains, or languages. Some key issues that may arise include how to design individual prompts for different tasks, and how to modulate their interaction with each other. So far this field has not been explored. Fig.5 illustrates a simple multiple prompt learning strategy for multiple tasks, where prompt templates are partially shared.

10.5 Selection of Pre-trained Models

With plenty of pre-trained LMs to select from (see §3), how to choose them to better leverage prompt-based learning is an interesting and difficult problem. Although we have conceptually introduced (§3.4) how different paradigms of pre-trained models are selected for diverse NLP tasks, there are few to no systematic comparisons of the benefits brought by prompt-based learning for different pre-trained LMs.

10.6 Theoretical and Empirical Analysis of Prompting

Despite their success in many scenarios, theoretical analysis and guarantees for prompt-based learning are scarce. Wei et al. (2021) showed that soft-prompt tuning can relax the non-degeneracy assumptions (the generation probability of each token is linearly independent) needed for downstream recovery (i.e. recover the ground-truth labels of the downstream task.), making it easier to extract task-specific information. Saunshi et al. (2021) verified that text classification tasks can be reformulated as sentence completion tasks, thus making language modeling a meaningful pre-training task. Scao and Rush (2021) empirically show that prompting is often worth 100s of data points on average across classification tasks.

10.7 Transferability of Prompts

Understanding the extent to which prompts are specific to the model and improving the transferability of prompts are also important topics. (Perez et al., 2021) show that prompts selected under tuned few-shot learning scenario (where one has a larger validation set to choose prompts) generalize well across models of similar sizes while prompts selected under true few-shot learning scenario (where one only has a few training samples) do not generalize as effectively as the former setting among models with similar sizes. The transferability is poor when the model sizes are quite different in both scenarios.

10.8 Combination of Different Paradigms

Notably, much of the success of the prompting paradigm is built on top of pre-trained models that were developed for the pre-train and fine-tune paradigm, such as BERT. However, are the pre-training methods that are effective for the latter applicable as-is to the former, or can we entirely re-think our pre-training methods to further improve accuracy or ease of applicability to prompting-based learning? This is an important research question that has not been covered extensively by the literature.

10.9 Calibration of Prompting Methods

Calibration (Gleser, 1996) refers to the ability of a model to make good probabilistic predictions. When using the generation probability of the pre-trained LMs (e.g., BART) to predict the answer, we need to be careful since the probability distribution is typically not well calibrated. Jiang et al. (2020b) observed the probabilities of pre-trained models (e.g., BART, T5, GPT-2) on QA tasks are well calibrated. Zhao et al. (2021) identify three pitfalls (majority label bias, recency bias and common token bias) that lead the pre-trained LMs to be biased toward certain answers when provided answered prompts. For example, if the final answered prompt has a positive label, then this will bias the model towards predicting positive words. To overcome those pitfalls, Zhao et al. (2021) first use context-free input (e.g. the prompt would be "Input: Subpar acting. Sentiment: Negative\n Input: Beautiful film. Sentiment: Positive\n Input: N/A. Sentiment:") to get the initial probability distribution P_0, then they use the real input (e.g. the prompt would be "Input: Subpar acting. Sentiment: Negative\n Input: Beautiful film. Sentiment: Positive\n Input: Amazing. Sentiment:") to get the probability distribution P_1. Finally, these two distributions can be used to get a calibrated generation probability distribution. However, this method has two drawbacks: (1) it comes with the overhead of finding proper context-free input (e.g. whether to use "N/A" or "None") and (2) the probability distribution of the underlying pre-trained LM is still not calibrated.

Even though we have a calibrated probability distribution, we also need to be careful when we assume a single gold answer for an input. This is because that all surface forms of a same object will compete for finite probability mass (Holtzman et al., 2021). For example, if we consider the gold answer to be "Whirlpool bath", the generation probability of it will typically be low since the word "Bathtub" shares the same meaning and it will take over a large probability mass. To address this issue, we could either (i) perform answer engineering to construct a comprehensive gold answer set using paraphrasing methods (§5.2.2) or (ii) calibrate the probability of a word based on its prior likelihood within the context (Holtzman et al., 2021).

11 Meta Analysis

In this section, we aim to give a quantitative birds-eye view of existing research on prompting methods by performing a meta analysis over existing research works along different dimensions.

TABLE 12 Timeline of prompt-based learning. The time for each paper is based on its first arXiv version (if exists) or estimated submission time. A web-version can refer to NLPedia-Pretrain. Works in red consider natural language understanding (NLU) tasks; works in blue consider natural language generation (NLG) tasks; works in green consider both NLU tasks and NLG tasks.

2018.06.07	LMComm (Trinh and Le, 2018)
2019.02.14	GPT-2 (Radford et al., 2019)
2019.04.14	WNLaMPro (Schick and Schütze, 2020)
2019.07.31	LMDiagnose (Ettinger, 2020)
2019.08.20	AdvTrigger (Wallace et al., 2019a)
2019.09.02	CohRank (Davison et al., 2019)
2019.09.03	LAMA (Petroni et al., 2019)
2019.09.11	CTRL (Keskar et al., 2019)
2019.10.23	T5 (Raffel et al., 2020)
2019.11.08	Neg & Misprim (Kassner and Schütze, 2020)
2019.11.28	LPAQA (Jiang et al., 2020c)
2019.12.10	ZSC (Puri and Catanzaro, 2019)
2020.01.21	PET-TC (Schick and Schütze, 2021a)
2020.03.10	ContxFP (Petroni et al., 2020)
2020.05.02	UnifiedQA (Khashabi et al., 2020)
2020.05.22	RAG (Lewis et al., 2020b)
2020.05.28	GPT-3 (Brown et al., 2020)
2020.09.08	CommS2S(Yang et al., 2020)
2020.09.15	PET-SGLUE (Schick and Schütze, 2021b)
2020.09.24	ToxicityPrompts (Gehman et al., 2020)
2020.10.07	WhyLM (Saunshi et al., 2021)
2020.10.13	X-FACTR (Jiang et al., 2020a)
2020.10.26	Petal (Schick et al., 2020)
2020.10.29	AutoPrompt (Shin et al., 2020)
2020.12.08	CTRLsum (He et al., 2020a)
2020.12.22	PET-Gen (Schick and Schütze, 2020)
2020.12.31	LM-BFF (Gao et al., 2021)
2021.01.01	WARP (Hambardzumyan et al., 2021)
2021.01.01	Prefix-Tuning (Li and Liang, 2021)
2021.01.17	KATE (Liu et al., 2021a)
2021.02.15	PromptProg (Reynolds and McDonell, 2021)
2021.02.19	ContxCalibrate (Zhao et al., 2021)
2021.02.24	PADA (Ben-David et al., 2021)
2021.02.27	SD (Schick et al., 2021)
2021.03.09	BERTese (Haviv et al., 2021)
2021.03.15	Prompt2Data (Scao and Rush, 2021)
2021.03.18	P-Tuning (Liu et al., 2021b)
2021.03.18	GLM (Du et al., 2021)
2021.03.22	ADAPET (Tam et al., 2021)
2021.04.10	Meta (Zhong et al., 2021a)
2021.04.12	OptiPrompt (Zhong et al., 2021b)
2021.04.14	Soft (Qin and Eisner, 2021)
2021.04.15	DINO (Schick and Schütze, 2021)
2021.04.15	AdaPrompt (Chen et al., 2021b)
2021.04.16	PMI_{DC} (Holtzman et al., 2021)
2021.04.18	Prompt-Tuning (Lester et al., 2021)
2021.04.18	Natural-Instr (Mishra et al., 2021)
2021.04.18	OrderEntropy (Lu et al., 2021)
2021.04.18	FewshotSemp (Shin et al., 2021)
2021.04.26	PanGu-α (Zeng et al., 2021)
2021.05.24	TrueFewshot (Perez et al., 2021)
2021.05.24	PTR (Han et al., 2021)
2021.06.03	TemplateNER (Cui et al., 2021)
2021.06.03	PERO (Kumar and Talukdar, 2021)
2021.06.16	PromptAnalysis (Wei et al., 2021)
2021.06.20	CPM-2 (Zhang et al., 2021)
2021.06.21	BARTScore (Yuan et al., 2021b)
2021.06.24	NullPrompt (Logan IV et al., 2021)
2021.06.25	Frozen (Tsimpoukelli et al., 2021)
2021.07.05	ERNIE-B3 (Sun et al., 2021)
2021.07.07	Codex (Chen et al., 2021a)
2021.07.14	HTLM (Aghajanyan et al., 2021)
2021.07.15	FLEX (Bragg et al., 2021)

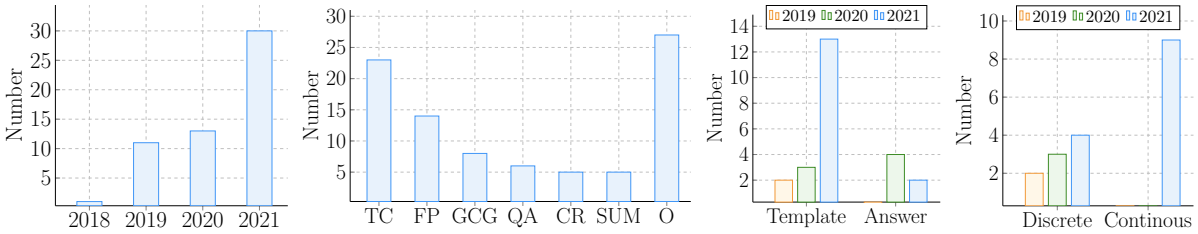
11.1 Timeline

We first summarize a number of existing research papers in a chronological order with in the form of a *timeline*, which hopefully, help researchers who are new to this topic understand the evolution of the field.

11.2 Trend Analysis

We also calculate the number of prompt-based papers with respect to different dimensions.

Year With the emergence of different kinds of pre-trained LMs, prompt-based learning has become a more and more active research field, as can be seen in Fig. 6-(a). We can see a huge surge in 2021, which is perhaps due to the prevalence of GPT-3 (Brown et al., 2020), which greatly increased the popularity of prompting in the few-shot multi-task setting.



(a) Year. (b) Task. (c) Automatic Search. (d) Search Space.

Figure 6: Meta-analyses over different dimensions. The statistics are based on the works in Tab. 7 and Tab. 8. In (d), we use the following abbreviations. TC: text classification, FP: factual probing, GCG: general conditional generation, QA: question answering, CR: commonsense reasoning, SUM: summarization, O: others.

Tasks We plot the number of works that investigate various tasks in Fig. 6-(b). For a task that has fewer than 5 relevant works, we group it into "Others". As the bar chart indicates, most tasks regarding prompt-based learning revolve around text classification and factual probing. We conjecture that this is because that for these tasks, both template engineering and answer engineering are relatively easy to conduct, and experiments are relatively computationally inexpensive.

Prompt vs. Answer Search As noted in previous sections, both prompt and answer search are important tools to take advantage of pre-trained language models for many tasks. Current research mainly focuses on template search instead of answer search, as shown in Fig. 6-(c).

Likely reasons are: (1) For conditional generation tasks (e.g. summarization or translation), the gold references can be directly used as answer. Although there are many sequences that may share the same semantics, how to effectively conduct multi-reference learning in conditional text generation problems is non-trivial. (2) For classification tasks, most of the time, label words are relative easy to select using domain knowledge.

Discrete Search vs. Continuous Search Since there are only a few works focus on automatic answer search, we analyze the automatic template search. As time goes by, there has been a shift from discrete search to continuous search for prompt engineering, as shown in Fig. 6-(d). Likely reasons are: (1) discrete search is harder to optimize compared to continuous search, (2) soft prompts have greater representation ability.

12 Conclusion

In this paper, we have summarized and analyzed several paradigms in the development of statistical natural language processing techniques, and have argued that *prompt-based learning* is a promising new paradigm that may represent another major change in the way we look at NLP. First and foremost, we hope this survey will help researchers more effectively and comprehensively understand the paradigm of prompt-based learning, and grasp its core challenges so that more scientifically meaningful advances can be made in this field. In addition, looking all the way back to the summary of the four paradigms of NLP research presented in §1, we hope to highlight the commonalities and differences between them, making research on any of these paradigms more full-fledged, and potentially providing a catalyst to inspire work towards the next paradigm shift as well.

Acknowledgements

We would like to thank Chunting Zhou for her constructive comments on this work.

References

[1] Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. Htlm: Hyper-text pre-training and prompting of language models. *arXiv preprint arXiv:2107.06955*.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *CoRR*, abs/2012.09816.

[3] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[5] Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

[6] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

[8] Eyal Ben-David, Nadav Oved, and Roi Reichart. 2021. Pada: A prompt-based autoregressive approach for adaptation to unseen domains.

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

[10] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

[11] Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland. Association for Computational Linguistics.

[12] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. 2016. Learning feed-forward one-shot learners. In *Advances in neural information processing systems*, pages 523–531.

[13] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.

[14] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

[15] Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. FLEX: unifying evaluation for few-shot NLP. *CoRR*, abs/2107.07170.

[16] 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. *arXiv preprint arXiv:2005.14165*.

[17] Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2017. Ask the right questions: Active question reformulation with reinforcement learning. *arXiv preprint arXiv:1705.07830*.

[18] Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

[19] Duo Chai, Wei Wu, Qinghong Han, Fei Wu, and Jiwei Li. 2020. Description based text classification with reinforcement learning. In *International Conference on Machine Learning*, pages 1371–1382. PMLR.

[20] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harri Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

[21] Xiang Chen, Xin Xie, Ningyu Zhang, Jiahuan Yan, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021b. Adaprompt: Adaptive prompt-based finetuning for relation extraction. *CoRR*, abs/2104.07650.

[22] Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021a. mt6: Multilingual pretrained text-to-text transformer with translation pairs. *CoRR*, abs/2104.08692.

[23] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021b. XLM-E: cross-lingual language model pre-training via ELECTRA. *CoRR*, abs/2106.16138.

[24] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

[25] J. Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555.

[26] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[27] Ronan Collobert, J. Weston, L. Bottou, Michael Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

[28] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

[29] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using bart.

[30] Hal Daumé III and Eric Brill. 2004. Web search intent induction via automatic query reformulation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 49–52, Boston, Massachusetts, USA. Association for Computational Linguistics.

[31] Joe Davison, Joshua Feldman, and Alexander M. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1173–1178. Association for Computational Linguistics.

[32] 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.

[33] Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

[34] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

[35] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

[36] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

[37] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All nlp tasks are generation tasks: A general pretraining framework.

[38] Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. 2011. Generalized minimum bayes risk system combination. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1356–1360.

[39] Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguistics*, 8:34–48.

[40] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

[41] Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

[42] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017a. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

[43] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017b. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

[44] Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. *arXiv preprint arXiv:2106.00641*.

[45] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1811–1820.

[46] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*.

[47] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3356–3369.

[48] Leon Jay Gleser. 1996. Measurement, regression, and calibration.

[49] Joshua T Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

[50] David Grangier and Michael Auli. 2018. QuickEdit: Editing text & translations by crossing words out. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 272–282, New Orleans, Louisiana. Association for Computational Linguistics.

[51] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

[52] Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

[53] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

[54] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

[55] Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. Warp: Word-level adversarial reprogramming. *ArXiv*, abs/2101.00121.

[56] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification.

[57] Ahmed Hassan. 2013. Identifying web search query reformulation using concept based matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1000–1010, Seattle, Washington, USA. Association for Computational Linguistics.

[58] 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.

[59] Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Fatema Rajani, and Caiming Xiong. 2020a. Ctrlsum: Towards generic controllable text summarization. *CoRR*, abs/2012.04281.

[60] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020b. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

[61] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[62] Ari Holtzman, Peter West, Vered Schwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right.

[63] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

[64] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

[65] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080.

[66] 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.

[67] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2020b. How can we know when language models know? *CoRR*, abs/2012.00955.

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

[69] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics.

[70] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

[71] Daniel Jurafsky and James H Martin. 2021. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[72] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*.

[73] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

[74] 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, ACL 2020, Online, July 5-10, 2020*, pages 7811–7818. Association for Computational Linguistics.

[75] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

[76] 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.

[77] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

[78] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

[79] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

[80] Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

[81] Sotiris B Kotsiantis, I Zaharakis, P Pintelas, et al. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

[82] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387. PMLR.

[83] Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning.

[84] J. Lafferty, A. McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

[85] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

[86] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

[87] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[88] Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

[89] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

[90] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

[91] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning.

[92] Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

[93] Hector J. Levesque. 2011. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

[94] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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.

[95] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[96] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

[97] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

[98] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

[99] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017a. Attention correctness in neural image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[100] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3?

[101] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3093–3102, Osaka, Japan. The COLING 2016 Organizing Committee.

[102] Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017b. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798.

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

[104] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

[105] 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.

[106] Yixin Liu, Zi-Yi Dou, and Pengfei Liu. 2021c. RefSum: Refactoring neural summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1448, Online. Association for Computational Linguistics.

[107] Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*.

[108] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020b. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4513–4519. ijcai.org.

[109] Robert L. Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models.

[110] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

[111] Yao Lu, Max Bartolo, A. Moore, S. Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *ArXiv*, abs/2104.08786.

[112] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

[113] Andreĭ Andreevich Markov. 2006. An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(4):591–600.

[114] Yvette Mathieu and Paul Sabatier. 1986. INTERFACILE: Linguistic coverage and query reformulation. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

[115] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

[116] Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

[117] Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

[118] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

[119] Tomas Mikolov, Ilya Sutskever, Kai Chen, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

[120] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *CoRR*, abs/2104.08773.

[121] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380, Florence, Italy. Association for Computational Linguistics.

[122] Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

[123] Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. *arXiv preprint arXiv:1704.04572*.

[124] Rodrigo Frassetto Nogueira, Jannis Bulian, and Massimiliano Ciaramita. 2019. Multi-agent query reformulation: Challenges and the role of diversity. *ICLR Workshop on Deep Reinforcement Learning for Structured Prediction*.

[125] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168, Boston, Massachusetts, USA. Association for Computational Linguistics.

[126] Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-M: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *CoRR*, abs/2012.15674.

[127] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

[128] Jeffrey Pennington, R. Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

[129] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models.

[130] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[131] Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

[132] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *ArXiv*, abs/2005.04611.

[133] 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.

[134] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

[135] Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *CoRR*, abs/1912.10165.

[136] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[137] 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.

[138] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

[139] Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. In *arXiv*.

[140] Alec Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *arXiv*.

[141] 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(140):1–67.

[142] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

[143] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

[144] Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. 2017. Learning to compose domain-specific transformations for data augmentation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3236–3246.

[145] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

[146] Itsumi Saito, Kyosuke Nishida, Kosuke Nishida, and Junji Tomita. 2020. Abstractive summarization with combination of pre-trained sequence-to-sequence and saliency models. *arXiv preprint arXiv:2003.13028*.

[147] Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2021. A mathematical exploration of why language models help solve downstream tasks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[148] Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493*.

[149] Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5569–5578. International Committee on Computational Linguistics.

[150] Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8766–8774. AAAI Press.

[151] Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.

[152] Timo Schick and Hinrich Schütze. 2020. Few-shot text generation with pattern-exploiting training.

[153] Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze questions for few shot text classification and natural language inference.

[154] Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners.

[155] Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp.

[156] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

[157] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

[158] Richard Shin, C. H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, D. Klein, J. Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. *ArXiv*, abs/2104.08768.

[159] 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 *Empirical Methods in Natural Language Processing (EMNLP)*.

[160] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

[161] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

[162] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

[163] Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

[164] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[165] Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.

[166] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019a. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

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

[168] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975. AAAI Press.

[169] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

[170] Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training.

[171] Kai Ming Ting and Ian H. Witten. 1997. Stacked generalizations: When does it work? In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes*, pages 866–873. Morgan Kaufmann.

[172] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

[173] Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

[174] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *CoRR*, abs/2106.13884.

[175] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. A wrong answer or a wrong question? an intricate relationship between question reformulation and answer selection in conversational question answering. In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, pages 7–16, Online. Association for Computational Linguistics.

[176] 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 *Advances in neural information processing systems*, pages 5998–6008.

[177] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2153–2162. Association for Computational Linguistics.

[178] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019b. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

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

[180] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34.

[181] Colin Wei, Sang Michael Xie, and Tengyu Ma. 2021. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning.

[182] Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

[183] Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1702–1715. Association for Computational Linguistics.

[184] Han Xu, Zhang Zhengyan, Ding Ning, Gu Yuxian, Liu Xiao, Huo Yuqi, Qiu Jiezhong, Zhang Liang, Han Wentao, Huang Minlie, et al. 2021. Pre-trained models: Past, present and future. *arXiv preprint arXiv:2106.07139*.

[185] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *CoRR*, abs/2105.13626.

[186] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

[187] Jheng-Hong Yang, Sheng-Chieh Lin, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Designing templates for eliciting commonsense knowledge from pretrained sequence-to-sequence models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3449–3453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[188] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

[189] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8413–8426. Association for Computational Linguistics.

[190] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. Association for Computational Linguistics.

[191] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*.

[192] Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021a. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

[193] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021b. Bartscore: Evaluating generated text as text generation.

[194] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation.

[195] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

[196] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020b. Human gaze assisted artificial intelligence: a review. In *IJCAI: Proceedings of the Conference*, volume 2020, page 4951. NIH Public Access.

[197] Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA. Association for Computational Linguistics.

[198] Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021. CPM-2: large-scale cost-effective pre-trained language models. *CoRR*, abs/2106.10715.

[199] 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

[200] Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, YuSheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020c. CPM: A large-scale generative chinese pre-trained language model. *CoRR*, abs/2012.00413.

[201] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

[202] Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021a. Meta-tuning language models to answer prompts better. *arXiv preprint arXiv:2104.04670*.

[203] Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021b. Factual probing is [MASK]: learning vs. learning to recall. *CoRR*, abs/2104.05240.

[204] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1-2):239–263.

[205] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Enhancing factual consistency of abstractive summarization. *arXiv preprint arXiv:2003.08612*.

[206] Geoffrey Zweig, John C. Platt, Christopher Meek, Christopher J.C. Burges, Ainur Yessenalina, and Qiang Liu. 2012. Computational approaches to sentence completion. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610, Jeju Island, Korea. Association for Computational Linguistics.

A Appendix on Pre-trained LMs

In this appendix we present some auxiliary information on pre-trained LMs that may be useful to the readers to better understand the current lay of the land with respect to this dynamic research area.

A.1 Evolution of Pre-trained LM Parameters

Fig. 7 lists several popular pre-trained models' statistics of parameters, ranging from 0 to 200 billion. GPT3, CPM2, and PanGu-α are the top three largest models with parameters greater than 150 billion.

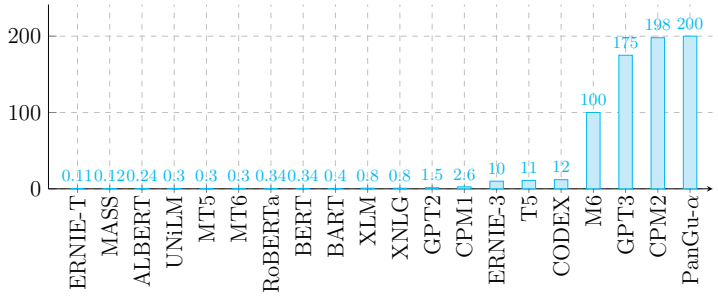


Figure 7: Comparison of the size of existing popular pre-trained language models.

A.2 Auxiliary Objective

In this subsection, more auxiliary objectives for pre-training language models have been listed.

- **Next Sentence Prediction (NSP)** (Devlin et al., 2019): A binary classification loss predicting whether two segments appear consecutively within a larger document, or are random unrelated sentences.
- **Sentence Order Prediction (SOP)** (Lan et al., 2020): A binary classification loss for predicting whether two sentences are in a natural or swapped order.
- **Capital Word Prediction (CWP)** (Liu et al., 2020b): A binary classification objective calculated over each word, predicting whether whether each word is capitalized or not.
- **Sentence Deshuffling (SDS)** (Liu et al., 2020b): A multi-class classification task to reorganize permuted segments.
- **Sentence distance prediction (SDP)** (Liu et al., 2020b) : A three-class classification task, predicting the positional relationship between two sentences (adjacent in the same document, not adjacent but in the same document, in different documents).
- **Masked Column Prediction (MCP)** (Yin et al., 2020): Given a table, recover the names and data types of masked columns.
- **Linguistic-Visual Alignment (LVA)** (Lu et al., 2019): A binary classification to Predict whether the text content can be aligned to visual content.
- **Image Region prediction (IRP)** (Su et al., 2020): Given an image whose partial features are masked (zeroed out), predict the masked regions.
- **Replaced Token Detection (RTD)** (Xiao et al., 2021): A binary classification loss predicting whether each token in corrupted input was replaced by a generative sample or not.
- **Discourse Relation Prediction (DRP)** (Sun et al., 2020): Predict the semantic or rhetorical relation between two sentences.
- **Translation Language Modeling (TLM)** (Lample and Conneau, 2019): Consider parallel sentences and mask words randomly in both source and target sentences.
- **Information Retrieval Relevance (IRR)** (Sun et al., 2020): Predict the information retrieval relevance of two sentences.
- **Token-Passage Prediction (TPP)** (Liu et al., 2020b): Identify the keywords of a passage appearing in the segment.
- **Universal Knowledge-Text Prediction (UKTP)** (Sun et al., 2021): Incorporate knowledge into one pre-trained language model.
- **Machine Translation (MT)** (Chi et al., 2021a) : Translate a sentence from the source language into the target language.
- **Translation Pair Span Corruption (TPSC)** (Chi et al., 2021a) : Predict the masked spans from a translation pair.
- **Translation Span Corruption (TSC)** (Chi et al., 2021a) : Unlike TPSC, TSC only masks and predicts the spans in one language.

- **Multilingual Replaced Token Detection (MRTD)** (Chi et al., 2021b): Distinguish real input tokens from corrupted multilingual sentences by a Generative Adversarial Network, where both the generator and the discriminator are shared across languages.
- **Translation Replaced Token Detection (TRTD)** (Chi et al., 2021b): Distinguish the real tokens and masked tokens in the translation pair by the Generative Adversarial Network.
- **Knowledge Embedding (KE)** (Wang et al., 2021): Encode entities and relations in knowledge graphs (KGs) as distributed representations
- **Image-to-text transfer (ITT)** (Wang et al., 2021): Is similar to the image caption that generates a corresponding description for the input image.
- **Multimodality-to-text transfer (MTT)** (Wang et al., 2021): Generate the target text based on both the visual information and the noised linguistic information.

A.3 Pre-trained Language Model Families

The increasing number of models makes it difficult for people to clearly grasp the differences between them. Based on this, we cluster the current mainstream pre-training models and characterize them from diverse dimensions.

Family	Models	LM	Pre-training Tasks			Corruption				Application
			Main	Auxiliary	Parallel	Mask	Replace	Delete	Permute	
GPT	GPT [139]	L2R	SLM	-	✗	-	-	-	-	NLG
	GPT-2 [140]	L2R	SLM	-	✗	-	-	-	-	NLG
	GPT-3 [16]	L2R	SLM	-	✗	-	-	-	-	NLG
	Codex [20]	L2R	SLM	-	✗	-	-	-	-	NLG
ELMo	ELMo [130]	L2R	SLM	-	✗	-	-	-	-	NLU, NLG
BERT	BERT [32]	Mask	CTR	NSP	✗	Tok	-	-	-	NLU
	RoBERTa [105]	Mask	CTR	-	✗	Tok	-	-	-	NLU
	SpanBERT [70]	Mask	CTR	-	✗	Span	-	-	-	NLU
	DeBERTa [60]	Mask	CTR	-	✗	Tok	-	-	-	NLU
	SciBERT [7]	Mask	CTR	NSP	✗	Tok	-	-	-	Sci-NLU
	BioBERT [89]	Mask	CTR	NSP	✗	Tok	-	-	-	Bio-NLU
	ALBERT [87]	Mask	CTR	SOP	✗	Tok	-	-	-	mSent
	FinBERT [108]	Mask	CTR	CWP, SDS, SDP, TPP	✗	Span	-	-	Sent	Fin-NLU
	VLBERT [164]	Mask	CTR	IRP	✓	Tok, Region	-	-	-	VLU
	ViLBERT [110]	Mask	CTR	IRP, LVA	✓	Tok, Region	-	-	-	VLU
	BEIT [5]	Mask	CTR,FTR	-	✗	Visual "Tok"[7]	-	-	-	VLU
	VideoBERT [166]	Mask	CTR	LVA	✓	Tok, Frame	-	-	-	VLU
	TaBERT [189]	Mask	CTR	MCP	✓	Tok, Column	-	-	-	Tab2Text
	mBERT [32]	Mask	CTR	NSP	✗	Tok	-	-	-	XLU
	TinyBERT [69]	Mask	CTR	NSP	✗	Tok	-	-	-	XLU
ERNIE	ERNIE-T [199]	Mask	CTR	NSP	✗	Tok, Entity	-	-	-	NLU
	ERNIE-B [169]	Mask	CTR	-	✗	Tok,Entity, Phrase	-	-	-	NLU
	ERNIE-NG [183]	Mask	CTR	RTD	✗	N-gram	Tok	-	-	NLU
	ERNIE-B2 [168]	Mask	CTR	CWP,SDS,SOP, SDP,DRP,IRR	✗	Entity, Phrase	-	-	Sent	NLU
	ERNIE-M [126]	LPM	CTR	-	✓	Tok	-	-		XLU, XLG
	ERNIE-B3 [167]	Mask	CTR	SOP,SDP,UKTP	✗	Entity, Phrase	-	-	-	NLU
BART	BART [94]	En-De	FTR	-	✗	Tok	Span	Tok	Sent,Doc	NLU, NLG
	mBART [104]	En-De	FTR	-	✗	Span	-	-	Sent	NLG
UniLM	UniLM1 [35]	LPM	SLM,CTR	NSP	✗	Tok	-	-	-	NLU, NLG
	UniLM2 [6]	LPM	SLM,CTR	-	✗	Tok	-	-	Tok	NLU, NLG
T5	T5 [141]	En-De	CTR	-	✗	-	Span	-	-	NLU, NLG
	mT5 [186]	En-De	CTR	-	✗	-	Span	-	-	XLU, XLG
	mT6 [22]	En-De	CTR	MT,TPSC,TSC	✓	-	Span	-	-	XLU, XLG
	ByT5 [185]	En-De	CTR	-	✗	-	byte-span	-	-	XLU, XLG
XLM	XLM [86]	LPM	CTR	TLM	✓	Tok	-	-	-	XLU, XLG
	XLM-R [28]	Mask	CTR	-	✗	Tok	-	-	-	XLU
	XLM-E [23]	Mask	CTR	MRTD,TRTD	✗	-	Tok	-	-	XLU, XLG
CPM	CPM [200]	L2R	SLM	-	✗	-	-	-	-	NLG
	CPM-2 [198]	En-De	CTR	-	✗	Span	-	-	-	NLU,NLG
Other	XLNet [188]	L2R	SLM	-	✗	-	-	-	Tok	NLU
	PanGu-α [194]	L2R	SLM	-	✗	-	-	-	-	NLG
	ELECTRA [26]	Mask	CTR	RTD	✗	Tok	Tok	-	-	NLU,NLG
	MASS [162]	En-De	CTR	-	✗	Span	-	-	-	NLG
	PEGASUS [195]	En-De	CTR	-	✗	Tok, Sent	-	-	-	Summarization
	M6 [179]	En-De	CTR	ITT,MTT	✗	Span	-	-	-	NLG

Table 13: A detailed illustration of different pre-trained models characterized by the four aspects. "**Parallel**" represents if parallel data have been used for pre-training. `Sci`, `Bio`, `Fin`, `K` represent scientific, biomedical, financial, and knowledge, respectively. `Tok`, `Sent`, `Doc` denote token, sentence and document, respectively. `Region`, `Frame` denote basic units of images and video respectively.