CSL: A Large-scale Chinese Scientific Literature Dataset for Cross-task Evaluation

Anonymous ACL submission

Abstract

Scientific literature serves as a high-quality corpus, which could provide natural annotated data for many natural language processing (NLP) research. In this work, we introduce a **C**hinese **S**cientific **L**iterature dataset – CSL, which contains the titles, abstracts, keywords and academic fields of 400,000 papers. The rich semantic information in these scientific literature creates extensive NLP tasks and provides a natural cross-task scenario. Based on this, we present a cross-task few-shot benchmark. To evaluate the cross-task transferability of the model, we design scenarios with different aspects and difficulties. Compared with previous cross-task benchmarks, these tasks are constructed from homogeneous corpus, allowing researchers to investigate the relationships between tasks, without being disturbed by heterogeneous data sources, annotation, and other factors. We analyze the behavior of existing text-to-text models on the proposed benchmark, and reveal the challenges for cross-task generalization, which provides a valuable reference for future research. Code and data are publicly available at GitHub[1].

1 Introduction

As the embodiment of human research knowledge, scientific literature is known as a rich source of informative data, supporting various NLP research (Luan et al., 2018; Cohan et al., 2019). So far, several scientific-related resources e.g. large-scale literature corpus (Lo et al., 2020; Saier and Färber, 2020), citation graphs (Sinha et al., 2015; Tang et al., 2008; Zhang et al., 2019), scientific downstream tasks (Lee et al., 2020; Beltagy et al., 2019) are available. Previous works, however, have primarily relied on digital libraries, such as arXiv, PubMed, CiteSeerX and ACL Anthology, which are mostly centered around the English language and focus on specific research fields.

To fill the gap of non-English scientific corpora, in this paper we introduce CSL: a large-scale **C**hinese **S**cientific **L**iterature dataset. CSL is obtained from 1982 Chinese core journals and contains meta-information of 400,000 papers with a wide range of distribution and fine-grained discipline annotation (67 categories).

Scientific literature metadata contains massive corpus information, making it a natural annotated data source with the potential to provide many high-quality NLP tasks. For example, predicting the title with abstract constitutes a summarization task, and predicting the discipline is a classification task. There are hundreds of such combinations. These tasks are constructed with homogeneous data, encouraging models to share knowledge across tasks.

Cross-task generalization, i.e., how to learn a new task efficiently based on the experiences of previous tasks, is an hot area in NLP community (Ye et al., 2021; Bragg et al., 2021; Sanh et al., 2021; Zhong et al., 2021). Previous studies mostly rely on heterogeneous data to create cross-task scenarios. For example, Ye et al. (2021) use 160 diverse NLP datasets to build a few-shot NLP gym; Bragg et al. (2021) use 20 dataset to construct transfer scenarios. For those cross-task scenarios above, there are multiple task-agnostic variables, such as data sources, annotation and task formats, making it difficult to reveal the relationship between specific tasks. In this paper, we introduce our cross-task benchmark, which includes a series of tasks where underlying knowledge and distribution are shared. We aim to reduce the variance of heterogeneous data and focus on evaluating connections among tasks, as well as providing a testbed for cross-task research.

We present three cross-task scenarios that are common in real applications. Each is made up of meta training tasks (meta tasks) and disjoint few-shot evaluation tasks (few-shot tasks). These scenarios show different relationships between tasks,

[1] https://github.com/CSL-Dataset/CSL_Dataset

Title	Abstract	Keywords	Discipline	Category
城市道路绿化景观研究 Research on Urban Road Greening Landscape	随着我国城市化步伐的加快... With the progress of urbanization in China ...	道路; 景观; 绿化; Road; Landscape; Greening;	园艺学 Horticulture	农学 Agriculture
分布式库存管理 Distributed Inventory Management Strategy	分析了分布式库存的管理模型... This paper analyzes the management model of distributed inventory ...	分布式库存; 协调中心; Distributed inventory; Coordination center;	应用经济学 Applied Economics	经济学 Economics

Table 1: Examples of the CSL dataset.

and vary in difficulties. It allows us to better understand how the connection of tasks affects the cross-task performance of the model. We provide a prompt-based text-to-text method as our baseline, which allows full parameter sharing across different task formats and easier transfer learning. Experiment results show that text-to-text language models are capable of cross-task transfer. However, in some challenging scenarios, there is still room for improvement.

The main contributions of this paper are summarized as follows:

- We release the first large-scale Chinese Scientific Literature dataset (CSL), which can be used for many different purposes, e.g. pre-training corpus and scientific-related tasks.
- Based on the CSL, we introduce a benchmark including different scenarios for cross-task few-shot evaluation.
- We propose a prompt-based method as our baseline and the experiment results highlight the model's difficulties in learning across tasks.

2 The CSL Dataset

2.1 Data Collection and Processing

We collect Chinese papers' homepage from publicly available search engines (Wanfang Data [2] and CNKI [3]) dated 2010 to 2020, then use the XSS parser to extract the meta-information in each web page, such as the title, abstract, keyword, and the journal in which the paper was published. To improve data quality, we filter the data as follows: First, we exclude papers that are not published in the Core Journals of China. Second, we filter out papers from comprehensive journals, and only preserve papers from professional journals.

According to Classification of Chinese Instructional Programs, academic fields are divided into 13 categories (Engineering, Medicine, etc.) and 67 disciplines (Mechanical Engineering, Oral Medicine, etc.). For each professional journal in the Core Journal Catalog, we associate it with a discipline based on the journal's description and published papers. Therefore, papers are annotated with two classification labels according to the published journal. For example, papers from "Chinese Journal of Computers" are categorized into Engineering category and Computer Science discipline.

Category	d	len(t)	len(a)	num(k)	#
Engineering	27	19.1	210.9	4.4	177k
Science	9	20.7	254.4	4.3	35k
Agriculture	7	17.1	177.1	7.1	39k
Medicine	5	20.7	269.5	4.7	37k
Management	4	18.7	157.7	6.2	23k
Jurisprudence	4	18.9	174.4	6.1	21k
Pedagogy	3	17.7	179.4	4.3	16k
Economics	2	19.5	177.2	4.5	11k
Literature	2	18.8	158.2	8.3	10k
Art	1	17.8	170.8	5.4	5k
History	1	17.6	181.0	6.0	6k
Strategics	1	17.5	169.3	4.0	3k
Philosophy	1	18.0	176.5	8.0	7k
All	67				400k

Table 2: The statistics of CSL dataset.

Finally, we collect 400K instances for CSL dataset, represented as a tuple $< t, a, k, c, d >$, where t is the title, a is the abstract, k is a list of keywords, c is the textual category label and d is the textual discipline label. CSL covers a wild range of academic fields, Table 1 shows the concrete examples in CSL dataset and the detailed statistics are provided in Table 2.

2.2 Task Formats

The current version of the CSL dataset contains 5 columns, which constitutes $\sum_{i=1}^{5}(\sum_{j=1}^{5-i} C_5^j \times C_{5-j}^i) = 180$ different tasks, where i and j are the number of input and output fields.

Previous works (Raffel et al., 2020; Gao et al., 2020) try to unify different tasks into general format for easier transfer learning. We extend this idea by designing task-specific prompts which cast

[2]https://www.wanfangdata.com.cn
[3]https://www.cnki.net

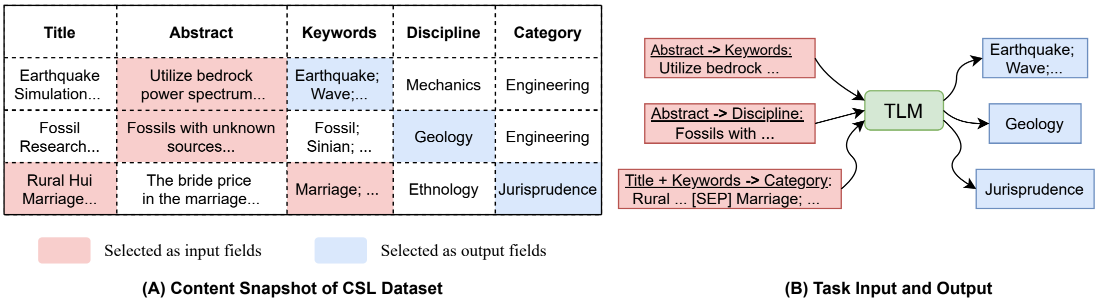

Figure 1: Overview of CSL task specific-prompts. **(A)** A content snapshot of the tabular CSL dataset. **(B)** TLM is a text-to-text language model. Prompts are added in front of input text such that different tasks can share loss function and output layers.

all CSL tasks into "text-to-text" format. For the CSL prompt, the input and output relationships are indicated by arrow characters, and multiple fields are coupled by plus signs. This prompt allows the model to predict multiple targets (i.e. one/many-to-many tasks) in a unified manner. Fig 1 gives a schematic overview of CSL prompts and tasks.

3 Cross-task Benchmark

Based on scientific NLP tasks derived form CSL dataset, we construct different cross-task scenarios, each of which contains pairs of meta tasks T_{meta} and disjoint few-shot test tasks T_{few}.

The model initially only has access to T_{meta} for meta training purpose, which captures the pattern in which the task structure differs from target tasks. The model's task-transferability is then evaluated by investigating the meta training stage's relative performance gain on learning T_{few}.

To comprehensively analyze the cross-task generalization, we manually design three scenarios concentrated on different aspects and varied in difficulties (i.e. the difficulty for few-shot tasks to leverage from meta tasks), each of which contains several partitions of T_{meta} and T_{few}. Tasks and their prompts are shown in Table 3. We sample 100 instances for each meta tasks and k-shot (k samples per class for classification tasks) for few-shot tasks with 8 different random seeds. 0, 1, 2, 4, 8 shot(s) are used for zero/few-shot training and 64 for validation/test.

Part.	Meta Tasks	Few-shot Tasks
Scenario 1		
1-1	Abst.→Kw., Kw→Title.	Abst.→Title.
1-2	Kw.→Title, Title→Dcp.	Kw.→Dcp.
1-3	Abst.→Title+Ctg., Title→Kw.	Abst.+Dcp.→Kw.
Scenario 2		
2-1	Abst.→Kw., Kw.→Title Title→Ctg.	Abst. →Ctg.
2-2	Abst.→Title., Title→Kw. Kw.→Dcp.	Abst. →Dcp.
Scenario 3		
3-1	Kw.→Abst., Title→Dcp.	Kw.→Dcp.
3-2	Abst.→Kw., Title→Kw. Title→Ctg.	Abst. →Ctg.

Table 3: Task prompts for each partition, Ctg: Category, Dcp: Discipline, Kw: Keywords, Abst: Abstract.

3.1 Single-leap Bridging

In *Scenario 1*, we evaluate the implicit bridging proposed by Johnson et al. (2017) as a zero-shot translation solution. It represents a real-world application where there is rare training data between the source field and the target field, an intermediate field can be utilized as a bridge. For example $Abstract \rightarrow Keywords$ and $Keywords \rightarrow Discipline$ enable a zero-shot task $Abstract \rightarrow Discipline$. Based on this, we design *partition*$_{1\text{-}1}$ and *partition*$_{1\text{-}2}$ representing naive single-leap implicit bridging tasks, *partition*$_{1\text{-}3}$ explores semi-connected bridging with one-to-many and many-to-one tasks.

3.2 Two-leap Bridging

Expanding the single-leap bridging, in *Scenario 2*, there are two intermediate fields between the source field and the target field, implying that three meta tasks are involved in the bridging. Additional meta tasks bring more samples for meta training, which theoretically have the same potential upper bound as *Scenario 1*. However, achieving it requires stronger transferability, which poses a tougher challenge for few-shot learners. In *partition*$_{2\text{-}1}$ and *partition*$_{2\text{-}2}$, we design meta tasks spanning through two leaps.

3.3 Broken Bridge

In real-world scenarios, it can be difficult to locate the intermediate field that connects sources and

#	Model	Scenario 1							Scenario 2						Scenario 3			
		$TS_{1\text{-}1}$		$CLS_{1\text{-}2}$		$KG_{1\text{-}3}$		Avg. Δ_m	$CLS_{2\text{-}1}$			$CLS_{2\text{-}2}$			$CLS_{3\text{-}1}$		$CLS_{3\text{-}2}$	
		few	+meta	few	+meta	few	+meta		few	+meta	Δ_m	few	+meta	Δ_m	+meta	Δ_m	+meta	Δ_m
1-shot	**T5$_{base}$**	9.3	28.8	1.5	3.9	1.1	14.4	+11.7	7.9	17.0	+18.8	1.8	5.3	+3.5	4.2	+2.7	15.2	+7.3
	BART$_{base}$	21.9	29.9	6.3	16.9	2.8	26.1	+14.0	31.1	40.0	+8.9	11.9	10.7	-1.2	16.4	+10.1	29.1	-2.0
2-shot	**T5$_{base}$**	11.7	28.1	1.6	4.5	2.0	16.7	+11.4	11.4	26.7	+5.6	2.6	3.3	+0.7	3.7	+2.2	23.7	+12.2
	BART$_{base}$	21.9	30.1	7.0	16.9	3.2	28.6	+14.5	31.2	43.3	+12.1	10.1	9.8	-0.4	16.2	+9.2	34.5	+3.3
8-shot	**T5$_{base}$**	17.6	30.5	1.6	7.1	9.7	13.8	+7.5	38.1	40.3	+2.2	6.8	12.7	+5.9	6.7	+5.1	30.3	-7.8
	BART$_{base}$	26.6	31.3	17.4	20.9	22.6	30.8	+5.4	49.7	47.7	-2.0	22.3	19.4	-2.9	20.0	+2.6	30.1	-4.3

Table 4: Performance of text-to-text models on cross-task evaluation. The columns **few** presents directly fine-tuning on T_{few}, the **+meta** is first fine-tuning on T_{meta} (meta-tuning) and then on T_{few}. Δ_m is the Average Relative Gain of meta-tuning. We report the average metrics for each task: TS: text summarization, KG: keyword generation, CLS: classification.

targets. In *Scenario 3*, we cut off the bridges in meta tasks to imitate this condition, making the few-shot tasks even more challenging. It aims to study the impact of non-bridging homologous meta tasks on few-shot target tasks. Based on *partition$_{1\text{-}2}$* and *partition$_{2\text{-}1}$*, we create *partition$_{3\text{-}1}$* and *partition$_{3\text{-}2}$* by modifying one of the meta tasks, which causes the bridge disconnected while the others unchanged.

4 Experiments

4.1 Baseline and Metrics

We consider different pre-trained text-to-text models including T5 (Raffel et al., 2020) and BART (Lewis et al., 2019) as our baselines. However, since there are few publically available Chinese versions of them, we conduct pre-training from scratch. Pre-training details are shown in Appendix A.

Following (Ye et al., 2021), we use multi-task fine-tuning to evaluate the above pre-trained models with and without meta tasks (i.e. meta-tuning) separately in each partition. Fine-tuning hyperparameters and other details are shown in Appendix B. We report the average of results over repeated experiments. The evaluation toolkit can be found in Appendix C.

For evaluation metrics, we adopt ROUGE-L and BLEU for summarization tasks; F1 and Bpref (Buckley and Voorhees, 2004) for keyword generation tasks. Classification tasks adopt accuracy and F1 macro as metrics. All the metrics are calculated at Chinese character level.

4.2 Results and Analysis

We observe from Table 4 that, the gain of meta-tuning (Δ_m) is positive on average, meaning that meta tasks generally improve task generalization. In Scenario 1, the meta tasks boost the performance of the few-shot tasks dramatically, and this benefit is sustainable as the shots increases.

For Scenario 2, the benefit of meta-tuning to few-shot tasks is first noticeable, but it quickly fades with the increment of training samples and eventually drops. Finally, the models fine-tuned directly on T_{few} sometimes outperform the meta-tuned models. For example, *partition$_{2\text{-}2}$* has the same target as *partition$_{1\text{-}2}$*, i.e. discipline classification, but it receives more informative input. As a result, it outperforms *partition$_{1\text{-}2}$* in direct few-shot fine-tuning. However, compared with the former, it gains less from meta-tuning for the average Δ_m drops by %4.9. This demonstrates that two-leap bridging indeed increases the difficulty.

In *Scenario 3*, we found the Δ_m of *partition$_{3\text{-}1}$* and *partition$_{3\text{-}2}$* show the similar changing trend as *partition$_{1\text{-}2}$* and *partition$_{2\text{-}1}$*. However, in comparison, they have decreased by 2.0% and 5.1% on average. The results suggest that meta-tuning on homologous tasks generally improves few-shot learning, and implicit that bridging is a key factor affecting the task generalization.

More detailed experiment results and our other findings are demonstrated in Appendix D.

5 Conclusion

In this paper, we provide a large-scale Chinese Scientific Literature dataset (CSL) and use it to evaluate few-shot cross-task generalization. This represents the challenge of addressing low-resource tasks with high-resource tasks. From the experiment results, we observe that homogeneous meta tasks generally improve few-shot learners, however, when the bridge is broken, this benefit becomes negligible. We also release an open-source toolkit for extensive evaluation.

References

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.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *arXiv preprint arXiv:2107.07170*.

Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596.

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.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

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

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, 125(3):3085–3108.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *arXiv preprint arXiv:2003.01355*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*.

Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019. Oag: Toward linking large-scale heterogeneous entity graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2585–2595.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhe Zhao, Hui Chen, Jinbin Zhang, Wayne Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training 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): System Demonstrations*, pages 241–246.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878.

A Pre-training Chinese Text-to-text Language Models

For pre-training Chinese text-to-text models, we follow the architecture, optimization, and hyperparameter choices described in (Raffel et al., 2020; Lewis et al., 2019; Zhang et al., 2020). Following Chinese BERT model (Devlin et al., 2019), we use the tokenizer with a vocabulary of 21,128 Chinese characters. Models are trained for 1,000,000 steps with the sequence length of 128 and then trained for 250,000 additional steps with 512 sequence length on CLUE Corpus (Xu et al., 2020) with the batch size of 512.

B Fine-tuning Hyperparameters

For comprehensive evaluation, we additional evaluate different sizes of T5, BART and PEGASUS. When fine-tuning different pre-trained models. we use the same hyperparameters. The settings of hyperparameters are as follows. The learning rate is set to 3e-4 for T5 and 1e-5 for BART and PEGASUS, the batch size is 32 for training on T_{meta} and 1 for training on T_{few}. We set the number of epochs to 15 with early stopping. The maximal input and output length are set to 256, which can be shortened according to the length of the task data to speed up training. All results are reported with greedy decoding (i.e. choosing the highest-probability logit at every timestep). All experiments are conducted on 1 Tesla V100 GPU and the results are collected over average of 8 episodes. Altogether, the experiments take around 2000 GPU hours, the full results are shown in Table 9.

C Pre-training and Evaluation Toolkit

We use UER-py[4] (Zhao et al., 2019) as our pre-training and fine-tuning platform. Based on which we implement a toolkit for cross-task evaluation including function modules: (1) Sampling K-shot examples for meta/few-shot tasks according CSL evaluation protocol. (2) Conducting meta-tuning on meta tasks and fine-tuning on few-shot tasks. Code, dataset and pre-trained models are available at `https://github.com/CSL-Dataset/CSL_Dataset`.

D Additional Analysis and Samples

We provide full results broken down by partitions in Table 9. In this section, we describe other findings.

Larger models perform better on few-shot tasks but small models are more likely to benefit from meta-training. As with previous common sense, larger models perform better on average for various tasks. However, smaller models have considerable task transferability, sometimes better than large models. For example, the average Δ_m of BART-base is 3.7%, while of BART-small is 4.4%.

Sentence-level denoising pre-training performs better on few-shot learning. In the majority of cases, BART and PEGASUS outperform T5. We speculate that it is because T5 is pre-trained with token-level denoising tasks, whereas BART and PEGASUS are pre-trained with sentence-level denoising tasks, which makes pre-training tasks closer to downstream tasks, allowing for easier few-shot transfer.

Meta-tuning on bridging tasks enables zero-shot learning. In *0-shot* rows, we present results of direct evaluating meta-tuned models on T_{few} without few-shot training, which indicates prompt-based text-to-text models have zero-shot task generalization.

Large-scale fine-tuning enables multi-task language generator. We fine-tuned a T5-base model with full CSL training set and all tasks. Output samples of summarization and keyword generation tasks are shown in Table 5 and Table 6. Conditional text generation samples are shown in Table 7. For the rules of constructing CSL prompt is extensible, which makes prompts feasible at the inference phase even it is unseen during the training. For example, "$Abstract \rightarrow Abstract$" indicates rewriting a given input abstract. Table 8 presents samples of zero-shot text rewriting task. The samples for these different tasks above all come from the same model.

[4]https://github.com/dbiir/UER-py

Prompt: 摘要->标题
Abstract -> Title
Input Text: 综述了纳米磁性粒子和磁性脂质体的制备方法, 同时简要介绍了磁性脂质体在磁性分离、靶向药物、热疗、组织工程和造影剂等领域的应用进展.
This paper briefly discusses magnetic nanoparticles and magnetic liposomes preparation methods, as well as the progress of magnetic liposomes in the fields of magnetic separation, targeted drugs, hyperthermia, tissue engineering, and contrast agents.
Prediction: 纳米磁性粒子和磁性脂质体的制备及应用进展
Progress in the Preparation and Application of Magnetic Manoparticles and Magnetic Liposomes
Ground Truth: 磁性脂质体的制备及应用研究进展
Progress in the Preparation and Application of Magnetic Liposomes

Prompt: 摘要->标题
Abstract -> Title
Input Text: 通过对美国职业排球运动员进行非结构性访谈研究美国职业排球运动员对赞助商和赞助行为的态度... 赞助商应尊重运动员的情感和观点,从而使双方都能获得长远利益.
Through interviews, research was conducted on the attitudes of American professional volleyball players regarding sponsors and sponsorship activities ... Sponsors should respect athletes' feelings and opinions in order for both sides to profit in the long run.
Prediction: 美国职业排球运动员对赞助商和赞助行为的态度研究
Research on American Professional Volleyball Players' Attitudes Towards Sponsors and Sponsorship Behaviors
Ground Truth: 美国排球运动员对赞助的态度分析
Analysis of American Volleyball Players' Attitudes towards Sponsorship

Table 5: Samples of summarization.

Prompt: 摘要->关键词
Abstract -> Keywords
Input Text: 位置跟踪是移动机器人自主导航中的一个主要任务...文中给出了一个基于线性系统模型的位置估计方法
Position tracking is a major task in the autonomous navigation of mobile robots ... This paper proposes a position estimation method based on the linear system model.
Prediction: 移动机器人; 位置估计; 线性系统模型; 线性近似
Mobile robots; Position estimation; Linear system model; Linear approximation
Ground Truth: 移动机器人; 位置跟踪; 卡尔曼滤波
Mobile robots; Position tracking; Kalman filter

Prompt: 摘要->关键词
Abstract -> Keywords
Input Text: 通过对祁连山自然保护区周边农牧民经济状况的调查发现阻碍经济发展的问题... 提出了发展生态旅游等适合本地区经济发展的模式.
Problems with economic development were discovered during an investigation of the economic conditions of farmers and herders in the Qilian Mountain Nature Reserve ... Ecotourism and other models for local economic development were proposed.
Prediction: 祁连山自然保护区; 农牧民; 经济发展模式
Qilian Mountain Nature Reserve; Peasants and herdsmen; Economic development model
Ground Truth: 祁连山自然保护区; 周边经济; 发展模式
Qilian Mountain Nature Reserve; Peripheral economy; Development model

Table 6: Samples of keywords generation.

Prompt: 标题->摘要
Title -> Abstract
Input Text: 藏东南地区生态旅游资源评价与开发建议
Evaluation and Development Suggestions on Ecotourism Resources in Southeastern Tibet
Prediction: 藏东南地区是我国重要的生态旅游资源区,其旅游资源的开发是一个重要的课题.本文从旅游资源的评价与开发的角度,对藏东南地区生态旅游资源开发的现状进行了分析,并提出了相应的开发建议.
Southeast Tibet is a significant ecotourism resource area in China, and the development of its tourism resources is a key issue. This paper focuses on the current of development of ecotourism resources in Southeast Tibet from the standpoint of tourism resource evaluation and development. Also, make recommendations for the future development.

Prompt: 标题->摘要
Title -> Abstract
Input Text: 促进我国开放式基金发展的七大对策
Seven Strategies to Promote the Development of China's Open Funds
Prediction: 文章分析了我国开放式基金所面临的困难,提出了目前我国开放式基金发展中存在的弊端,提出了对策:要加快开放式基金发展,探索新型开放式投资工具.
This paper highlighted the problems facing China's open-end funds, identifies flaws in the current development of open-end funds, and suggests solutions, including accelerating the development of open-end funds and exploring new open-end investment tools.

Table 7: Samples of conditional text generation.

Prompt: 摘要->摘要 Abstract -> Abstract **Input Text:** 自蔓延高温合成(shs)技术是一种材料制备的新工艺,但该技术存在的某些局限性阻碍了它的进一步发展和工业化应用. Self-propagating high-temperature synthesis (shs) technology is a new process for material preparation, but some limitations of this technology hinder its further development and industrial application. **Prediction:** 自蔓延高温合成方法是一种将非金属超细胞材料及聚醚基金属离子转换成金属基金属的新技术,它以其独特的特点而成为一种化学研究的热点.该技术在材料制备领域的应用受到科技企业的认可和研究,但它的优越性并没有显著影响其在工业上商业化应用的进一步发展. The self-propagating high-temperature synthesis method is a new technology that converts non-metallic supercellular materials and polyether-based metal ions into metal-based metals. With its unique characteristics, it has become a hot spot in chemical research. In the field of material preparation, it is recognized and researched by technology companies, but its superiority has not significantly affected the further development of its commercial applications in industry.

Prompt: 摘要->摘要 Abstract -> Abstract **Input Text:** 烟大轮渡铁路栈桥采用两跨式全焊变截面下承式钢板梁,跨度为52.645m+29.855m.从结构形式、提升方式等方面介绍该栈桥的十大关键性技术. The Yantai Ferry Railway Trestle uses a two-span fully welded variable cross-section through steel plate girder with a span of 52.645m+29.855m. Ten key technologies of the trestle are introduced from the aspects of structure, lifting methods, etc. **Prediction:** 烟大轮渡桥是烟西市第一座高龄全焊变截面下承式钢板梁斜拉桥.在整体展示桥梁结构、桥梁板结构等方面介绍了该桥的关键技术,进而为该桥长期保持桥梁的强度构造奠定了有利实践基础. Yanxi City's first all-welded variable cross-section steel plate girder cable-stayed bridge is the Yantai Ferry Bridge. This paper introduces the bridge's key technologies in the context of the overall display of the bridge structure and the bridge slab structure, and then lays a favorable practical foundation for the bridge's long-term strength.

Table 8: Samples of rewriting.

		Scenario 1													Scenario 2											Scenario 3					
		$TS_{1\text{-}1}$				$CLS_{1\text{-}2}$				$KG_{1\text{-}3}$					$CLS_{2\text{-}1}$					$CLS_{2\text{-}2}$					$CLS_{3\text{-}2}$			$CLS_{3\text{-}2}$		
	Models	few		+meta		few		+meta		few		+meta		Avg.	few		+meta			few		+meta			+meta			+meta		
		R-L	B-4	R-L	B-4	Acc	F1	Acc	F1	F1	Bpr.	F1	Bpr.	Δ_m	Acc	F1	Acc	F1	Δ_m	Acc	F1	Acc	F1	Δ_m	Acc	F1	Δ_m	Acc	F1	Δ_m
0-shot	$T5_{small}$			37.9	12.4			3.0	2.8			1.4	3.8				10.3	6.1				4.9	5.1							
	$T5_{small}$			14.1	1.7			12.9	11.4			0.8	0.8				16.7	15.3				2.4	1.8							
	$BART_{base}$			39.4	15.3			18.9	18.3			2.6	3.1				33.3	30.0				13.9	13.8							
	$BART_{small}$			36.5	11.1			19.4	19.1			2.3	4.2				34.6	25.4				12.1	11.6							
	$PEGA_{\text{-}base}$			37.0	10.9			23.9	21.3			1.6	3.1				37.2	30.9				15.7	15.4							
	Average			33.0	10.3			15.6	14.6			1.7	3.0				26.4	21.5				9.8	9.5							
1-shot	$T5_{base}$	16.9	1.6	40.4	17.1	2.0	1.0	4.3	3.5	0.1	2.0	5.9	22.9	+11.7	9.6	6.2	19.9	14.2	+9.1	2.3	1.2	5.6	5.0	+3.5	4.5	3.8	+2.7	17.3	13.1	+7.3
	$T5_{small}$	5.8	0.3	34.7	11.5	2.8	2.3	12.6	11.8	0.1	0.0	3.5	15.1	+13.0	10.9	9.5	27.2	23.7	+15.2	2.4	1.3	4.7	4.3	+2.6	4.7	4.5	+2.0	14.4	12.6	+3.3
	$BART_{base}$	31.1	12.7	40.1	19.7	7.8	4.9	18.2	15.6	1.6	4.0	15.6	36.7	+14.0	33.3	28.9	44.2	35.7	+8.9	13.7	10.1	12.3	9.0	-1.2	17.8	15.1	+10.1	33.3	24.9	-2.0
	$BART_{small}$	31.0	12.3	38.3	17.5	9.7	7.0	18.8	15.3	4.5	8.3	13.9	33.2	+10.7	36.9	31.3	47.1	38.4	+8.6	14.1	10.4	15.0	11.1	+0.8	16.9	14.7	+7.4	40.4	33.3	+2.7
	$PEGA_{\text{-}base}$	41.1	21.1	42.6	21.8	18.2	14.4	25.9	22.4	6.0	21.4	10.7	27.3	+4.7	45.5	41.6	51.3	45.4	+4.8	20.7	16.6	21.9	18.4	+1.5	23.3	20.1	+5.4	47.8	39.6	+0.2
	Average	25.2	9.6	39.2	17.5	8.1	5.9	16.0	13.7	2.5	7.1	9.9	27.0	+10.8	27.2	23.5	37.9	31.5	+9.3	10.6	7.9	11.9	9.6	+1.4	13.4	11.6	+5.5	30.6	24.7	+2.3
2-shot	$T5_{base}$	21.0	2.3	39.7	16.5	2.2	0.9	5.1	4.0	0.4	3.6	6.3	27.1	+11.4	12.8	10.1	29.2	24.2	+15.2	3.4	1.9	3.9	2.8	+0.7	4.0	3.4	+2.2	26.0	21.4	+12.2
	$T5_{small}$	9.1	0.6	31.5	10.9	3.0	1.9	8.4	7.4	0.2	0.1	3.1	17.0	+10.6	11.2	7.7	26.9	24.6	+16.3	4.0	2.5	7.8	6.3	+3.8	4.9	4.1	+2.0	15.4	13.3	+4.9
	$BART_{base}$	31.0	12.9	40.2	20.0	8.7	5.3	18.4	15.5	1.1	5.4	15.5	41.7	+14.5	35.8	26.6	47.4	39.2	+12.1	11.8	8.4	11.2	8.3	-0.4	17.3	15.1	+9.2	38.5	30.4	+3.3
	$BART_{small}$	31.8	12.3	40.2	19.7	10.0	6.5	18.1	15.2	3.8	7.8	12.9	33.3	+11.2	30.8	23.3	41.3	32.4	+9.8	13.0	9.4	12.4	7.6	-1.2	19.1	15.9	+9.2	41.0	33.2	+10.1
	$PEGA_{\text{-}base}$	42.6	20.2	42.3	22.1	16.5	12.5	27.2	23.4	7.3	22.4	13.4	33.8	+6.8	48.1	42.6	47.8	39.2	-1.9	19.8	15.2	18.5	15.6	-0.4	24.3	21.4	+8.4	47.4	39.9	-1.7
	Average	27.1	9.7	38.8	17.8	8.1	5.4	15.4	13.1	2.6	7.9	10.2	30.6	+10.9	27.7	22.1	38.5	31.9	+10.3	10.4	7.5	10.8	8.1	+0.5	13.9	12.0	+6.2	33.7	27.6	+5.8
4-shot	$T5_{base}$	26.5	3.4	43.2	19.8	5.2	3.9	5.1	4.3	0.3	5.6	4.6	21.9	+9.0	32.1	27.0	41.0	35.0	+8.4	4.8	3.8	15.7	13.8	+10.4	3.9	2.8	-1.2	42.6	36.1	+9.8
	$T5_{small}$	7.3	0.3	35.5	11.7	5.4	3.9	13.7	11.9	0.1	0.0	3.1	23.3	+13.7	20.5	19.5	36.9	31.0	+14.0	10.7	8.5	8.7	7.9	-1.3	6.6	6.0	+1.6	30.4	26.4	+8.4
	$BART_{base}$	31.3	13.5	40.6	20.2	22.1	17.8	22.6	19.1	5.8	21.0	14.8	47.7	+8.9	51.9	46.8	55.1	48.5	+2.5	25.6	21.3	22.6	18.6	-2.9	22.1	18.6	+0.4	50.0	44.4	-2.1
	$BART_{small}$	31.3	12.7	39.5	18.5	21.1	17.8	22.5	18.5	5.8	15.1	13.8	44.4	+8.9	52.2	46.7	56.1	49.6	+3.4	24.8	20.9	23.6	19.5	-1.3	22.3	20.0	+1.7	56.1	50.5	+3.8
	$PEGA_{\text{-}base}$	45.2	23.6	43.3	22.8	28.2	23.8	27.9	23.7	12.5	33.2	13.3	37.8	+0.4	55.4	49.7	61.5	54.5	+5.5	29.3	25.1	28.8	24.9	-0.4	25.4	21.8	-2.4	55.4	50.4	+0.4
	Average	28.3	10.7	40.4	18.6	16.4	13.4	18.4	15.5	4.9	15.0	9.9	35.0	+8.2	42.4	37.9	50.1	43.7	+6.8	19.0	15.9	19.9	16.9	+0.9	16.1	13.8	+0.0	46.9	41.6	+4.1
8-shot	$T5_{base}$	27.9	7.4	42.4	18.7	2.4	0.8	8.1	6.1	3.0	16.4	4.9	22.7	+7.5	40.7	35.5	43.9	36.7	+2.2	7.5	6.1	13.3	12.1	+5.9	8.2	5.2	+5.1	32.7	28.0	-7.8
	$T5_{small}$	14.8	1.2	31.8	11.8	4.0	2.4	10.9	9.2	0.2	0.2	3.5	26.1	+11.8	25.3	20.8	32.4	28.5	+7.4	11.6	9.6	13.7	12.4	+2.5	4.5	3.7	+0.9	23.7	18.7	-1.9
	$BART_{base}$	35.8	17.5	41.4	21.2	19.5	15.3	22.9	18.9	14.8	30.4	16.5	45.1	+5.4	52.6	46.8	52.2	43.1	-2.0	24.5	20.1	21.6	17.1	-2.9	21.2	18.7	+2.6	49.0	41.8	-4.3
	$BART_{small}$	36.6	16.9	41.8	21.1	21.0	16.5	23.4	19.3	10.2	19.8	12.8	45.4	+7.1	51.6	46.3	53.2	46.2	+0.8	26.7	22.5	22.2	18.8	-4.1	22.3	19.1	+2.0	47.8	40.9	-4.6
	$PEGA_{\text{-}base}$	44.5	23.1	43.9	23.2	26.4	22.1	31.3	26.1	15.4	34.3	13.3	42.3	+2.4	57.4	52.0	61.9	55.1	+3.8	31.7	27.7	27.6	23.1	-4.3	25.7	22.8	+0.0	58.7	51.9	+0.6
	Average	31.9	13.2	40.3	19.2	14.7	11.4	19.3	15.9	8.7	20.2	10.2	36.3	+6.8	45.5	40.3	48.7	41.9	+2.4	20.4	17.2	19.7	16.7	-0.6	16.4	13.9	+2.1	42.4	36.3	-3.6

Table 9: Full experiment results. R-L: ROUGE-L, B-4: BLEU-4gram, Bpref: Binary preference, Acc: Accuracy. Note that there is no **few** column for $CLS_{3\text{-}1}$ and $CLS_{3\text{-}2}$ because they use shared few-shot tasks.