CSL: A Large-scale Chinese Scientific Literature Dataset

Yudong Li[1,2], **Yuqing Zhang**[1*], **Zhe Zhao**[3], **Linlin Shen**[2], **Weijie Liu**[3],
Weiquan Mao[3], **and Hui Zhang**[4]

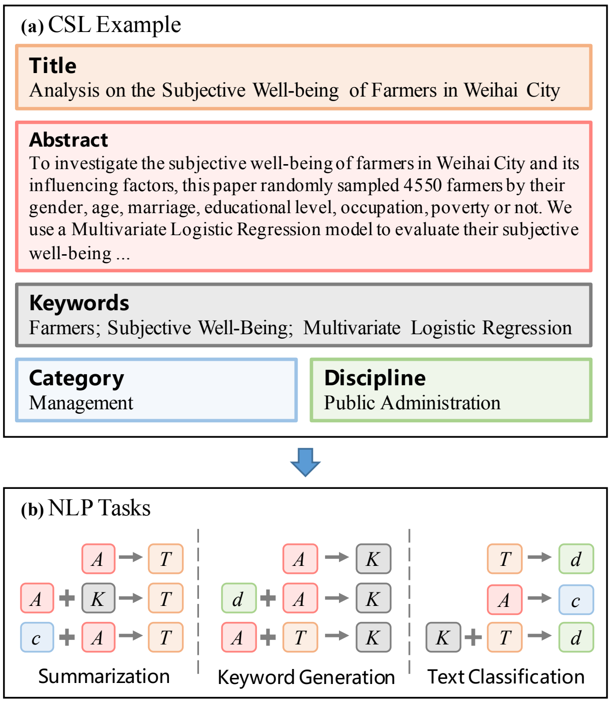
[1] China University of Geosciences (Beijing), School of Information Engineering
[2] Shenzhen University, School of Computer Science and Software Engineering
[3] Tencent AI Lab
[4] Information Technology Center for National Science
and Technology Infrastructure, Beijing, China

Abstract

Scientific literature serves as a high-quality corpus, supporting a lot of Natural Language Processing (NLP) research. However, existing datasets are centered around the English language, which restricts the development of Chinese scientific NLP. In this work, we present CSL, a large-scale **C**hinese **S**cientific **L**iterature dataset, which contains the titles, abstracts, keywords and academic fields of 396k papers. To our knowledge, CSL is the first scientific document dataset in Chinese. The CSL can serve as a Chinese corpus. Also, this semi-structured data is a natural annotation that can constitute many supervised NLP tasks. Based on CSL, we present a benchmark to evaluate the performance of models across scientific domain tasks, i.e., summarization, keyword generation and text classification. We analyze the behavior of existing text-to-text models on the evaluation tasks and reveal the challenges for Chinese scientific NLP tasks, which provides a valuable reference for future research. Data and code are available at `https://github.com/ydli-ai/CSL`.

1 Introduction

With the increase in the publication of papers, Natural Language Processing (NLP) tools that assist users in writing, searching, and archiving scientific literature have grown increasingly important. For instance, paper/citation recommendation (Beel et al., 2016; Cohan et al., 2020), topic classification (Beltagy et al., 2019) and summarization (Cohan and Goharian, 2018) systems have been developed. The construction of these automatic systems primarily relies on academic resources such as large-scale corpus (Lo et al., 2020; Saier and Färber, 2020), citation graphs (Sinha et al., 2015; Tang et al., 2008; Zhang et al., 2019) and supervised scientific datasets (Li et al., 2016; Jurgens et al., 2018). These resources, however, are mostly centered around the English language, which restricts the development of techniques for addressing non-English scientific NLP tasks. Until recently, progresses in the NLP research for Chinese resources and models has lagged behind their English counterparts.



Title
Analysis on the Subjective Well-being of Farmers in Weihai City

Abstract
To investigate the subjective well-being of farmers in Weihai City and its influencing factors, this paper randomly sampled 4550 farmers by their gender, age, marriage, educational level, occupation, poverty or not. We use a Multivariate Logistic Regression model to evaluate their subjective well-being ...

Keywords
Farmers; Subjective Well-Being; Multivariate Logistic Regression

Category
Management

Discipline
Public Administration

Figure 1: (a) An example of paper meta-information (translated into English). (b) Examples of NLP tasks constructed from CSL. The arrow indicates the input and output of the task, for example, "$A \rightarrow T$" represents the task feeding abstract to produce title. A: abstract; T: title; K: keywords; c: category; d: discipline.

To fill the gap of non-English scientific resources and spur the Chinese scientific NLP research, in this paper, we introduce CSL: a large-scale **C**hinese **S**cientific **L**iterature dataset. CSL contains 396,209 Chinese papers' meta-information, including title, abstract, keywords, academic category and discipline. Papers are collected from comprehensive Chinese academic journals covering a wide range of distribution. In particular, we divide them into

Proceedings of the 29th International Conference on Computational Linguistics, pages 3917–3923
October 12–17, 2022.

Category	#*d*	len(T)	len(A)	num(K)	#Samples	Discipline Examples
Engineering	27	19.1	210.9	4.4	177,600	*Mechanics, Architecture, Electrical Science*
Science	9	20.7	254.4	4.3	35,766	*Mathematics, Physics, Astronomy, Geography*
Agriculture	7	17.1	177.1	7.1	39,560	*Crop Science, Horticulture, Forestry*
Medicine	5	20.7	269.5	4.7	36,783	*Clinical Medicine, Dental Medicine, Pharmacy*
Management	4	18.7	157.7	6.2	23,630	*Business Management, Public Administration*
Jurisprudence	4	18.9	174.4	6.1	21,554	*Legal Science, Political Science, Sociology*
Pedagogy	3	17.7	179.4	4.3	16,720	*Pedagogy, Psychology, Physical Education*
Economics	2	19.5	177.2	4.5	11,558	*Theoretical Economics, Applied Economics*
Literature	2	18.8	158.2	8.3	10,501	*Chinese Literature, Journalism*
Art	1	17.8	170.8	5.4	5,201	*Art*
History	1	17.6	181.0	6.0	6,270	*History*
Strategics	1	17.5	169.3	4.0	3,555	*Military Science*
Philosophy	1	18.0	176.5	8.0	7,511	*Philosophy*
All	67				396,209	

Table 1: Detailed statistics of the CSL dataset. #d: The number of disciplines in the category. len(T): Average length of each title; len(A): Average length of each abstract; num(K): Average number of keywords.

13 first-level categories and 67 second-level disciplines. In addition to the difference in language, it provides broader and more fine-grained research fields than existing academic resources (Lo et al., 2020; Saier and Färber, 2019).

Scientific literature metadata contains abundant semantic information, making it a natural annotated data source with the potential to provide many NLP tasks. For example, predicting the title with abstract constitutes a summarization task. As the data and task examples shown in Figure 1, such combinations can constitute abundant tasks. These tasks can drive models in real-world applications and are essential for a lot of academic NLP research. To better understand the challenges posed by Chinese scientific NLP, we build a benchmark consisting of a series of CSL-derived tasks, i.e., summarization, keyword generation and category/discipline classification. We also provide a toolkit that allows users to design evaluation tasks according to their needs.

We implement some state-of-the-art Chinese text-to-text models and evaluate on the proposed benchmark. We also demonstrate the effectiveness of the CSL dataset as pre-training corpus. Specifically, we pre-train T5 with paper abstracts, namely CSL-T5. It outperforms the model trained on the general-domain corpus, which can be used as a strong baseline for the proposed benchmark. The experiment results show that though existing models can achieve acceptable performance on scientific NLP tasks, it still needs future efforts to reach a practical level.

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 build a benchmark that represents real-world scenarios of automatic analyzing scientific literature.
- We implement text-to-text models to provide baselines. The experimental results highlight the model's difficulties in Chinese scientific NLP tasks.

2 The CSL Dataset

2.1 Data Collection

We obtain the paper's meta-information from the National Engineering Research Center for Science and Technology Resources Sharing Service (NSTR) [1] dated from 2010 to 2020. Then, we filter data by the Catalogue of Chinese Core Journals, which is an academic journal evaluation standard published by Peking University Library. It selects 1,980 core journals from the Chinese journals, widely recognized by the Chinese academic community.

According to the Catalogue and collected data, we divide academic fields into 13 first-level categories (e.g., Engineering, Science) and 67 second-level disciplines (e.g., Mechanics, Mathematics). We use the journal's instructions to assign journals to categories and disciplines, and only journals that focus on a single academic field are kept. For the guideline of human annotation, we follow the Disciplines of Conferring Academic Degrees (GB/T 13745-2009). We ask volunteers to read the intro-

[1] https://nstr.escience.net.cn

Dataset	Instances	Language	Peer Review	Source	Academic Disciplines
S2ORC (2020)	8.1M	English	not all	MAG, arXiv, PubMed	20 (multi)
PubMed Central OAS	2.3M	English	not all	PubMed	2 (bio, LS)
unarXive (2020)	1.0M	English	not all	MAG, arXiv	4 (physics, math, CS, other)
Saier and Färber, 2019	1.0M	English	not all	arXiv	3 (physics, math, CS)
arXiv CS (2018)	90k	English	not all	arXiv	1 (CS)
AAN (2013)	25k	English	all	ACL Anthology	1 (comp ling)
CSL (ours)	396k	Chinese	all	Chinese Core Journals	67 (multi)

Table 2: A comparison of CSL with other publicly-available scientific literature datasets. Note that we provide the first dataset in Chinese, which also has the more fine-grained discipline annotation. bio: biomedicine; LS: life science; CS: computer science; comp ling: computational linguistics.

duction of the journal and find the closest discipline from the guideline. As a result, papers can be labeled with categories and disciplines based on the journal in which they were published. For example, papers from the "Chinese Journal of Computers" are categorized into the category "Engineering" and the discipline "Computer Science".

In total, we collect 396,209 instances for the CSL dataset, represented as tuples $< T, A, K, c, d >$, where T is the title, A is the abstract, K is a list of keywords, c is the category label and d is the discipline label. Due to the ethical concern, we only use the paper's publicly available meta-information and do not access the full text.

2.2 Data Analysis

The paper distribution over categories and the examples of disciplines are shown in Table 1. A total of 67 disciplines are collected by CSL, covering almost all research fields. Each discipline contains 3000-10000 samples.

Table 2 presents an overview of existing academic datasets. In comparison, the CSL dataset has the following features: **(1) Wider discipline coverage.** Existing datasets mainly focus on specific domains, while CSL covers almost all research domains. Also, CSL has more fine-grained discipline labels. **(2) New data source.** It can be seen that existing datasets are largely built on digital libraries like arXiv [2], PubMed [3], ACL Anthology [4] and MAG (Sinha et al., 2015), resulting in some overlap. CSL presents a new data source in Chinese that complements existing resources. **(3) Higher quality and accuracy.** Digital libraries contain pre-print platforms, and therefore some papers are not peer-reviewed. CSL is collected from published journal papers and is potentially of higher quality. In addition, CSL directly accesses the database without PDF/LaTeX parsing, which has near-perfect accuracy.

2.3 Evaluation Benchmark

The CSL contains meta-information provided by authors when submitting their papers, and the connections between them can constitute many NLP tasks. In this section, we build a benchmark to facilitate the development of Chinese scientific literature NLP. It contains diverse tasks, ranging from classification to text generation, representing many practical scenarios. We randomly select 10,000 samples and split the datasets into training set, validation set and test set according to the ratio, 0.8 : 0.1 : 0.1. This split is shared across different tasks, which allows multi-task training and evaluation. From CSL, many possible combinations can also constitute tasks. We provide a toolkit for users to design tasks by their needs.

Text Summarization (TS) The paper title can be seen as a summary of the paper abstract. We build a summarization task predicting the paper title from the abstract. Existing Chinese text summarization resources are mainly concentrated in the news domain (Hu et al., 2015; Liu et al., 2020), and we provide the first text summarization task in the academic domain.

Keyword Generation (KG) In this task, the model is asked to predict a list of keywords from a given paper title and abstract. This task is similar to the Paper Topic Classification (Cohan et al., 2020), but instead of predicting topics in a set of candidates, the goal is to generate keywords that correspond to the paper. We construct a dataset of paper's keywords, title and abstract. In English, there are related datasets such as SemEval (Kim et al., 2013) and KP20k (Meng et al., 2017). To

[2] https://arxiv.org

[3] http://www.pubmed.gov

[4] https://aclanthology.org

the best of our knowledge, CSL provides the first Chinese keyword generation task.

Text Classification This task is predicting the category and discipline based on other information about the paper. We build a dataset for **category classification (CTG$_{\text{cls}}$)**, which predicts the category with the paper title. Besides, we build a **discipline classification (DCP$_{\text{cls}}$)** task that predicts the discipline with the paper abstract.

3 Experiments

3.1 Baseline Models

For baselines, we evaluate multi-task learning models trained on CSL tasks. We use the text-to-text (i.e., feed text to produce text) method to unify downstream tasks in different formats. Specifically, these tasks are represented as the language generation task guided by a textual prompt. We adopt several widely used text-to-text models, including T5 (Raffel et al., 2020), PEGASUS (Zhang et al., 2020), and BART (Lewis et al., 2019). Since there are few publicly available versions of them, we conduct pre-training on the Chinese corpus from scratch. In addition, we train a T5 using CSL paper abstracts as the corpus, namely CSL-T5, to provide a pre-training model that adapts to the Chinese scientific domain.

3.2 Settings

For pre-training Chinese text-to-text models, we follow the architecture, optimization, and hyperparameter choices described in the papers. Following Google's Chinese BERT (Devlin et al., 2019), we use the tokenizer with a vocabulary of 21,128 Chinese characters. Models are pre-trained on the CLUE Corpus Small (Xu et al., 2020) for 1M steps with the batch size of 512. We progressively train CSL-T5 basis on pre-trained T5, using the paper abstract as the corpus for 20k steps with the same hyperparameters.

Experiments are conducted on UER-py framework (Zhao et al., 2019) [5].The learning rate is set to $3e^{-4}$ for T5; $1e^{-5}$ for BART and PEGASUS. The batch size is 32. For multi-task training, we combine the training sets of each task for training 5 epochs. We use a prompt to specify which task the model should perform, e.g., "to category" for category classification. Then, we fine-tune the models on the task to be evaluated for 3 epochs with early stopping. All results are reported with greedy decoding.

Models	CTG$_{\text{cls}}$	DCP$_{\text{cls}}$	TS	KG
	Acc.	Acc.	R-L	Bpref.
T5	**83.6**	67.1	49.8	54.2
PEGASUS	81.7	69.4	49.4	55.2
BART	79.2	65.7	47.8	49.9
T5 (single)	82.3	63.2	49.2	54.1
CSL-T5	82.9	**70.8**	**52.1**	**55.9**

Table 3: The test performances of baseline models on CSL downstream tasks. T5 (single) is the result of fine-tuning T5 on each task separately, and the remaining columns are the results of multi-task learning.

Prompt: to 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
Prompt: to 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 4: Samples of text summarization and keyword generation of CSL-T5.

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

3.3 Overall Performance

The experimental results are shown in Table 3. Output samples of text summarization and keyword generation tasks are shown in Table 4. For text classification, we report accuracy. We use ROUGE-L (Lin and Hovy, 2003) for the summarization task, which is commonly used for language generation tasks. For keyword generation, we use Bpref. (Buckley and Voorhees, 2004), which evaluates both the accuracy and order of generated keywords. We can observe that baseline models can achieve acceptable results, where T5 outperforms other models. However, it is still not satisfactory for real-world applications, and future efforts are needed. We also find that domain-adaptive training can further improve T5's performance. Similar experiments are also done by Beltagy et al. (2019) and Gururangan et al. (2020), it partially demonstrates the value of CSL corpus for pre-training. The model and corpus will be publicly available.

To discover the effect of multi-task training, we fine-tune T5 with each task individually. From the comparison between T5 and T5-single, multi-task learning slightly outperforms individually fine-tuned models. We speculate that since the CSL tasks are homogeneous (derived from the same dataset), it is easier for models to share knowledge across different tasks.

CSL can create a large number of tasks by different combinations of tasks' input and output. It provides a natural playground for observing which tasks are mutually reinforcing when learned together. CSL could also be useful for cross-task research (Ye et al., 2021; Bragg et al., 2021). For example, exploring which tasks the model learns can help it quickly adapt to new tasks. We leave that for future exploration.

4 Conclusion

This paper presents the first Chinese scientific literature dataset, CSL, which can serve as a pre-training corpus and can derive abundant NLP tasks. Based on CSL, we build an evaluation benchmark to explore the challenges posed by automatic analysis of Chinese scientific documents. Experimental results find difficulties in existing models in the Chinese scientific domain and point out the future directions.

Limitations and future work. In the current version of CSL, to provide accurate category/discipline labels, we only use journals focused on one domain, which resulted in some data loss. In future work, we will provide multi-label datasets to cover cross domain papers, and annotate CSL with more attributes like Chinese-English parallel data for academic machine translation. Also, the versatile NLP task derived from CSL constitutes a naturally cross-task scenario. In the future, we will explore the role of CSL in cross-task and few-shot research.

Ethical Considerations

The corpus we use is released by the Chinese government aimed at sharing academic resources, which has been anonymized wherever necessary. We are licensed to use some of paper's metadata for NLP research. Therefore, our dataset does not involve any privacy or copyright issues.

Acknowledgments

This research was supported by National Natural Science Foundation of China under grant no. 91959108 and Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515111199. We appreciate the data preprocessing and annotation effort done by Zichun Cao, Peixia Zhang, and Zhihui Wu in Prof. Zhang's lab at CUGB.

References

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338.

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. *Advances in Neural Information Processing Systems*, 34.

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, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings* *of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Arman Cohan and Nazli Goharian. 2018. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, 19(2):287–303.

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.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. A high-quality gold standard for citation-based tasks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2013. Automatic keyphrase extraction from scientific articles. *Language resources and evaluation*, 47(3):723–742.

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*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Xiaojun Liu, Chuang Zhang, Xiaojun Chen, Yanan Cao, and Jinpeng Li. 2020. Clts: A new chinese long text summarization dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 531–542. Springer.

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.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

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. 2019. Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)*, volume 2345, page 14–26.

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.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189.

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.