Research on Fine-grained S&T Entity Identification with Contextual Semantics in Think-Tank Text

Mengge Sun[1,2], Yanpeng Wang[1,2,*] and Yang Zhao[1,2]

[1]National Science Library, Chinese Academy of science Beijing 100190

[2]Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences Beijing 100190

Abstract

Automatically extracting fine-grained S&T problems from think-tank reports written by numerous experts, has become one of the effective ways to perceive the global trend of S&T development. We transform the automatic identification task for fine-grained S&T problems into a multi category S&T entity extraction task with contextual semantics. To address the shortage of high-quality data sets and fully exploit the potential of LLMs, we take LLMs as annotators and puts them into an active learning loop to determine which samples to annotate efficiently. During the cyclic data annotation process, we simultaneously trained the target's entity extraction model "RoBERTa-BiLSTM-CRF". Finally, the model achieved an F1 value of 86.02% in our task. The effectiveness and reliability of the model were verified by comparing it with the benchmark model through experiments. This study to some extent solves the problem of manually annotating dataset dependencies, while providing high-quality data support and effective model methods for mining and analyzing fine-grained S&T problems.

Keywords

S&T entity with contextual semantics, LLM annotators, active learning, RoBERTa-BiLSTM-CRF,

1. Introduction

The think tank is composed of multidisciplinary experts in a country and gathers national intellectual resources, which is an important force to influence government decision-making and promote social development. Usually, think tank reports tend to focus on major issues of great concern to the national government or the public, which represent indicators and weather vane of national policies and scientific research, and have high intelligence values. Therefore, the automatic extraction of scientific and technological problems mentioned in think tank reports can further clarify policy and public concerns efficiently and objectively. This paper defines "fine-grained S&T problems" as "research directions or problems with limited conditions such as application scenarios, technological solutions, and technological routes", and further analogizes them as "S&T entities with contextual semantics".

Most of the S&T problem representations extracted by researchers in the past have adopted several methods such as manual annotation, rule-based matching, machine learning-based, hybrid model-based, and deep learning-based methods. H. Chu and Q. Ke [1] used manual annotation to analyze the distribution of methods in different academic journals. However, those expert annotation methods are relatively highly accurate, but costly and time-consuming. S. Gupta and C D. Manning [2] designed matching rules for identifying research problem , including using the word "applied" for rule matching, and then using the Bootstrapping method to find new rule templates based on the newly matched vocabulary. K. Heffernan and S. Teufel [3] treated scientific method identification as a classification task, using classification algorithms such as support vector machines, Naive Bayes, and logistic regression, and introduced features such as N-gram, sentiment polarity, part of speech, whether it is a negative word, discourse information, and part of speech into the algorithm to enhance its performance. Semeval 2018 Task7 [4] also conducted extraction of various types of entities in academic papers. In this task, many teams used convolutional neural networks and Long Short-Term Memory networks to achieve performance superior to traditional machine learning methods (such as SVM), which also proved the usefulness of deep learning models. In terms of deep learning methods, Xuesi Li et al. [5] designed a sentence classification model based on the BERT-CNN architecture, and automatically identified research issue sentences in scientific papers with an F1 value of 94.8%. Z. Zhong and D. Chen [6] compared the performance of BERT and SciBERT, two pre-trained language models, in the extraction of relations in academic papers, and found that SciBERT performed better than BERT.

Since 2020, large language models (LLMs) have exhibited remarkable few-shot performance in information extraction tasks, with only a few demonstrations and well-designed prompts. Under the prevalent "Language-Model-as-a-Service" (Sun et al. 2022) setting, users are

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), April 23 24, 2024, Changchun, China and Online

*Corresponding author.

wangyanpeng@mail.las.ac.cn (Y. Wang)

required to feed their own data, potentially including sensitive or private information, which increases the risk of data leakage.To exploit the abundant unlabeled corpus, an alternative is to employ LLMs as annotators, which generate labels in a zero-shot or few-shot manner.

In this paper, we subdivide S&T entities into multiple grained categories. Depending on the type of scientific solution sought, they can be distinguished into: identification and judgment about the research object and the inherent mechanisms and laws of research. Correspondingly, the research objects include **"technological methods", "system devices", "scientific experiments", "scientific materials", and "databases name".** Examples include "cell-based cancer immunotherapy and gene therapy", "ferrosilicon alloy latent heat photovoltaic cells", "deep underground neutrino experiments" and "two-dimensional materials for future heterogeneous electronic devices". And the underlying mechanisms of things, such as "the principle of evolution controlled from top to bottom".

2. Data and Methods

2.1. Data

The selected data source is high-quality strategic dynamic briefing data monitored and compiled by various departments of the Chinese Academy of Sciences and the State Council, which is available on the agency's website[1] . The data source includes: (1) the trends of top scientific journals, showcasing the latest scientific research achievements in disciplines such as physics, Earth, and biology; (2) the latest strategic deployments of various countries in the field of S&T, representing the direction of national S&T development.

These information contents can to some extent represent the will of the country and scientists [7]. Finally, we crawled all the information from the three sites from 2018 to 2023, totaling 42,984 reports, with an average of about 12 sentences per report.

2.2. Main Framework

Based on the above data sets, the research work of this paper mainly includes three parts: initial annotation based on syntatic rules, active annotation based on LLM, and train extraction model during active learning process. As shown in figure 1.

[1]http://www.casisd.cn/zkcg/ydkb/kjqykb/
https://news.sciencenet.cn/AInews/newlist.aspx?
http://www.globaltechmap.com/document/index

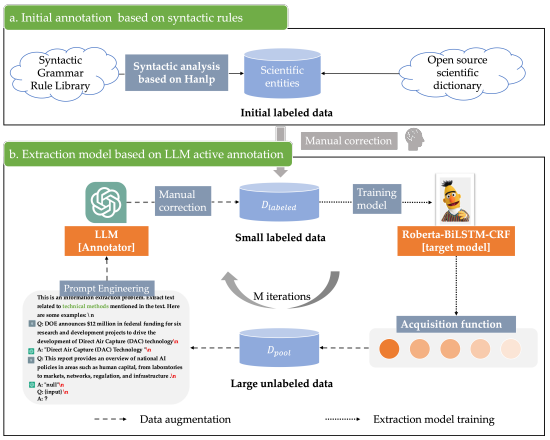

Figure 1: Main research framework

2.2.1. Initial annotation based on syntactic rules

In the part of initial annotation based on syntactic rules, we mainly uses a rule-based extraction method as a cold start, combined with manual correction, to obtain a small amount of high-quality contextual S&T entities databases. As of now, there are a total of 162 lexical and syntactic rules. Then, combined with the dependency syntax analysis function of a pretrained HanLP model, candidate scientific entity phrases with contextual semantics are obtained.

2.2.2. Extraction model based on LLM active annotation

In the extraction part of based on LLM active annotation, the main goal is to gradually fine screen a small-scale annotated data from a large amount of unlabeled data, while using a large language model as the annotation model. At the same time, a S&T entity extraction model called "Roberta-BiLSTM-CRF" is trained.

1) Optimizing LLM as better annotator. According to literature research, it has been found that the current GPT series models are highly sensitive to different PROMPT expressions. When different annotators use different PROMPT expressions, there is a significant difference in the response results of GPT. The robustness of the model on NLP tasks is relativaly weak [8]. Previous studies show that the design of task-specific prompts varies between near state-of-the-art and random guesses [9]. Therefore, finding the best prompts for given tasks and given data points is very critical.

This paper adopts the Chain of thought (CoT) prompts strategy, which gradually generates label sequences that meet expectations by setting some conditions in each model. Guided by the CoT approach, this article transforms this task into a multi round Q&A question, en-

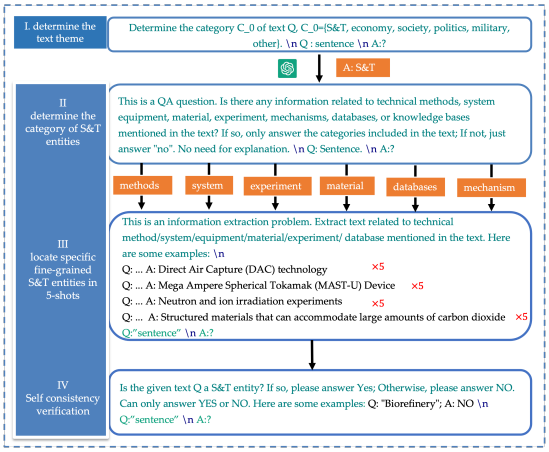

Figure 2: Flowchart of GPT annotation under CoT

abling the GPT model to gradually locate the fine-grained categories of S&T entities contained in the text through conversation, and finally annotate them. Specifically, this chapter focuses on the construction process of PROMOPT for different categories of S&T problems, as shown in Figure 2.

2) Active data acquisition. Active learning (AL) seeks to reduce labeling efforts by strategically choosing which examples to annotate. We consider the standard pool-based setting, assuming that a large pool of unlabeled data D_{pool} is available. AL loop starts with a seed labeled set $D_{labeled}$. **At each iteration, we train a model M on $D_{labeled}$ and then use acquisition function $f(\cdot, M)$ to acquire a batch B consisting of b examples from D_{pool}.** We then query the LLM annotator to label B. The labeled batch is then removed from the pool D_{pool} and added to labeled set $D_{labeled}$, and will serve as training data for the next iteration. The process is repeated for m times.

Active acquisition strategies generally maximize either uncertainty or diversity. On one hand, uncertainty-based methods(such as Maximum Entropy, Least Confidence) leverage model predictions to select hard examples. On the other hand, diversity-based methods(such as K-Means) exploit the *heterogeneity of sampled data.*

2.2.3. Extraction model based on Roberta-BiLSTM-CRF

Training the target model based on the labeled data obtained, and select the data to be annotated in the next iteration using the acquisition function mechanism. Among them, the target model uses the Chinese RoBERTa-WWM[10] model as the embedding model, and the BiLSTM model and CRF model as the label sequence prediction layer to obtain the label sequence of

Table 1
Performance of the Large Model at Each Prompt Stage in final iteration

ID	Precision	Recall Rate
Stage I	100.0	–
Stage II	90.87	–
Stage III	71.20	88.41
Stage IV	92.01	–

S&T entities and complete the automatic extraction of fine-grained S&T problem.Finally, evaluate the model results based on soft matching strategy.

3. Results and discussion

3.1. Analysis of data annotation results

Initial supervised data based on the rule annotation In the data annotation based on statistical rules, it was found that the extraction effect of the model was: accuracy: 0.36; recall rate: 0.82; F1 value: 0.50. That is to say, the majority of S&T phrases annotated by statistical rule-based annotation methods are not within the category of S&T, and their level of S&T cannot be accurately judged. Therefore, an AI model that can deeply understand and analyze semantics is particularly needed for annotation and extraction.

3.2. Analysis of LLM annotation results

Firstly, we randomly selected 20 texts from the annotated dataset as the test set to determine the number of examples in the Few-shot strategy.The results showed that 5-shot had the highest accuracy at 76.3%. The test shows that the more relevant and semantically similar the given examples are to the test text, the better the annotation effect of GPT3.5. In the 1-shot scenario where an example is given, the performance of the given example is sensitive and unstable to GPT3.5; Overall, 5-shot prompt performs better because combining multiple random examples can reduce the impact of noise.

After determining the number of given examples in the Few shot strategy, we conducted multiple tests to select the most effective example for each stage of the PROMPT. The performance of each prompt stage is shown in Table 1.

(1) In terms of category judgment, GPT's performance is almost perfect. That is to say,for classification tasks with more popular query semantics and more obvious semantic differences, the GPT model has better performance.

Table 2
Model comparison experiment results

METHOD	Precision	Recall Rate	F1 Value
PROMPTING	67.72	76.72	67.72
BERT-BiLSTM-CRF	70.54	75.66	73.00
Our model	**82.20**	**90.23**	**86.02**

(2) In terms of information extraction, GPT has lower accuracy and higher recall.The extracted S&T entities are mainly in the form of nouns phrases, which are not comprehensive, such as "natural language processing algorithms will be used to study the principle of virus gene mutation.

Finally, after multiple rounds of annotation and manual proofreading, a total of 19745 sentences formed a supervised training dataset.

3.3. Analysis of Model extraction effect

Dataset We chose 2680 fine-grained S&T entities datasets as seed labeled set $D_{labeled}$ from initially annotated dataset, use the whole 19745 sentences as D_{pool} and randomly acquired 100 samples per batch for 10 iterations, which generate 9,921 annoted samples $D_{labeled}$ in total.

Baselines We compare RoBERTa-BiLSTM-CRF with the following baselines: (1) In-context learning (i.e. PROMPTING). The PROMPTING enables LLM to conduct few-shot inference without fine-tuning. (2) SUPERVISED(i.e. BERT-BiLSTM-CRF). The surpervised model is trained on whole clean-labeled data $D_{labeled}$.

Accelerating with Active Learning The last layer in the above extraction model is the CRF model, whose output result is the probability score of the BIO label corresponding to each character. Here, we use this probability score as the confidence score and input it into two uncertainty based active learning strategies. The results show that maximal entropy active learning strategies enable extraction model to be more efficient and more capable. The results of the S&T entities extraction tasks are shown in Table 2.

4. Conclusions

Automatic extraction of contextual technology entities with contextual semantics from think tank reports can more efficiently capture key research development directions. In this paper, GPT is used as the teacher model and Roberta-Bilstm-CRF is used as the student model. Through active learning method, the training data generated by GPT is fine-tuned to the local extraction model, forming a set of feasible fine-grained S&T entity recognition framework.

The biggest limitation of our study is that it mainly focuses on the discussion of the effectiveness of the method, and the standard of high accuracy has not been reached in practical engineering applications, and the model effect will continue to be optimized in the future.

References

[1] H. Chu, Q. Ke, Research methods: What's in the name?, Library & Information Science Research 39 (2017) 284–294.

[2] S. Gupta, C. D. Manning, Analyzing the dynamics of research by extracting key aspects of scientific papers, in: Proceedings of 5th international joint conference on natural language processing, 2011, pp. 1–9.

[3] K. Heffernan, S. Teufel, Identifying problems and solutions in scientific text, Scientometrics 116 (2018) 1367–1382.

[4] D. Buscaldi, A.-K. Schumann, B. Qasemizadeh, H. Zargayouna, T. Charnois, Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers, in: International Workshop on Semantic Evaluation (SemEval-2018), 2017, pp. 679–688.

[5] Y. L. Y. W. Xuesi Li, Zhixiong Zhang, Research on problem sentence recognition methods in scientific literature research, Library and Information Service 67 (2023) 132–140.

[6] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, arXiv preprint arXiv:2010.12812 (2020).

[7] X. C. Y. L. X. L. Yanpeng Wang, Xuezhao Wang, Analysis of key technologies and initiatives of the fourth industrial revolution based on science and technology policy and frontier dynamics, Journal of the China Society for Scientific and Technical Information 41 (2022) 29–37.

[8] J. Gao, H. Zhao, C. Yu, R. Xu, Exploring the feasibility of chatgpt for event extraction, arXiv preprint arXiv:2303.03836 (2023).

[9] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, arXiv preprint arXiv:2012.15723 (2020).

[10] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, Pre-training with whole word masking for chinese bert, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3504–3514.