ITAKE: Interactive Unstructured Text Annotation and Knowledge Extraction System with LLMs and ModelOps

Jiahe Song [1,3] **Hongxin Ding**[1,3] **Zhiyuan Wang**[1,3] **Yongxin Xu**[1,3]
Junfeng Zhao[1,3,4*] **Yasha Wang**[1,2]

[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[2]National Engineering Research Center For Software Engineering, Peking University, Beijing, China
[3]School of Computer Science, Peking University, Beijing, China
[4]Big Data Technology Research Center, Nanhu Laboratory, Jiaxing, China
songjh@stu.pku.edu.cn, {zhaojf, wangyasha}@pku.edu.cn

Abstract

Extracting structured knowledge from unstructured text data has a wide range of application prospects, and a pervasive trend is to develop text annotation tools to help extraction. However, they often encounter issues such as single scenario usage, lack of effective human-machine collaboration, insufficient model supervision, and suboptimal utilization of Large Language Models (LLMs). We introduces an interactive unstructured text annotation and knowledge extraction system that synergistically integrates LLMs and ModelOps to alleviate these issues. The system leverages LLMs for enhanced performance in low-resource contexts, employs a ModelOps platform to monitor models throughout their lifecycle, and amalgamates interactive annotation methods with online machine learning and active learning. The demo video[1] and website[2] are now publicly available.

1 Introduction

Unstructured text data contains a large amount of valuable knowledge, from which structured knowledge such as entities, relationships and attributes can be extracted to help the construction of knowledge graphs, and can also support downstream tasks, which has a wide range of application prospects. However, real-world text exists multi-language, a mixture of short and long text, and complex terminological references, etc. Unstructured text knowledge extraction methods based solely on machine intelligence are far from meeting the needs of actual business. For example, on the publicly available datasets WNUT-17 (Derczynski et al., 2017), DocRED (Yao et al., 2019), the highest F1-score for named entity recognition and relation extraction are only 60.45% (Wang et al., 2021) and 67.53% (Ma et al., 2023). Besides, the cost of relying only on human annotation is very expensive.

Currently, there are many open-source text annotation tools dedicated to solving the above challenges, but they have some problems resulting in a not-so-perfect process. First of all, some of the tools are used in a single scenario, targeting a fixed application domain, ontology and language (**Challenge C1**). For example, MedCat (Kraljevic et al., 2021) only supports English and is limited to medical data annotation. Secondly, most of the tools lack the organic combination of human and machine, resulting in too much user participation to increase the cost (Stenetorp et al., 2012; Nakayama et al., 2018) or lack of user feedback leading to poor modeling accuracy (Zhang et al., 2022b) especially in low resource situation (**Challenge C2**). In addition, even if models are involved in the extraction process of some tools (Kraljevic et al., 2021; Zhang et al., 2022b), there is a lack of model supervision and state analysis in the process of using them, and the reuse support capability for models and datasets is weak, which prevents the rapid development and deployment of models for specific domain requirements (**Challenge C3**). Finally, after the popularity of LLMs (Brown et al., 2020; Touvron et al., 2023; Du et al., 2022) , many extraction tools intergrated LLMs to assist extraction (Wei et al., 2023; Zhang et al., 2022b). However, although LLMs are more effective than traditional knowledge extraction state-of-the-art model **(hereinafter referred to as the extraction model)** in low resource situation because of their strong generalization ability, the improvement effect of LLMs is not obvious after the increase of training data, and when they reaches a certain threshold, their effect is far worse than that of well-trained extraction model (Wang et al., 2023a). At the same time, LLMs are conversational generative models, which lead to slower inference speed and are difficult to meet the real-time demand (**Challenge C4**).

* Corresponding Author
[1]https://youtu.be/d_8vbdzdIe8
[2]http://itake.askgraph.site

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 326–334
August 11-16, 2024

Aiming at the above problem, we developed **ITAKE** (an **I**nteractive unstructured **T**ext **A**nnotation and **K**nowledge **E**xtraction system) that integrates LLMs and ModelOps (Hummer et al., 2019). Specifically, (1) addressing **Challenge C1** and **Challenge C3**, we adopt ModelOps platform to integrate different models and monitor whole lifecycle of them. (2) Addressing **Challenge C2**, we combine the interactive annotation methods for online machine learning (Fontenla-Romero et al., 2013) and active learning (Shen et al., 2017). (3) Addressing **Challenge C4**, we integrate LLMs under low resources situation and use extraction models for well-labeled situation.

2 Architecture

ITAKE consists of two subsystems as **Fig.1** shows.

2.1 Intelligent Knowledge Extraction Based on Human-Machine Collaboration Subsystem

This part consists of three parts: Project Management, Pre-annotation and Model Selection, Model Tuning and Batch Knowledge Extraction. **Project Management** is to manage the information and users of each knowledge extraction task; **Pre-annotation and Model Selection** is designed for domain experts to perform unsupervised knowledge extraction of unstructured data using LLMs; **Model Tuning and Batch Knowledge Extraction** uses active learning to selectively annotate fewer data in order to train the optimal model to the user's desired accuracy, after which it can proceed to batch knowledge extraction.

2.2 ModelOps-based Full Lifecycle Monitor of Models Subsystem

This part consists of five parts: LLMs Service (fine-tuning and extraction), Knowledge Extraction Model Selection and Recommendation Service, Knowledge Extraction Model Pool, Datasets Management and Model Lifecycle Management. Specifically, **LLMs Service** provides support for LLM fine-tuning such as ChatGLM (Du et al., 2022), Baichuan (Baichuan, 2023) and extraction, which solves the knowledge extraction cold start problem (Wang et al., 2023a); **Knowledge Extraction Model Selection and Recommendation Service** obtains the models from the model pool and performs training and comparison to provide the optimal models; **Knowledge Extraction Model Pool** accesses different models to solve the problems of nested entity and overlapped relationship, and unifies the management of a series of extraction models; **Model Lifecycle Management** unifies the release, management, and retrieval of LLMs and extraction models; **Datasets Management** can save and reuse knowledge extraction results.

3 Modules

3.1 Project Management

Project management encompasses tasks such as dataset uploading and data cleansing. ITAKE's upload interface supports different language texts, ontology models and file-type. Furthermore, ITAKE's backend deploys well-fine-tuned LLMs and well-trained extraction models for different domains, and by combining the above features, ITAKE can provide good extraction support for texts in different domains, thus solving the **Challenge C1**.

To ensure that the text datasets align with the requirements for subsequent knowledge extraction, ITAKE offers customizable rules for data cleansing and organization. Given the varied structure and content of unstructured text, datasets exhibit unique compositional features and semantic emphases. To address this, ITAKE introduces "iterative algorithms for user selection," empowering users to tackle these challenges effectively. The system is equipped with a range of universal algorithms at the backend, which can be dynamically invoked by users via the frontend interface, facilitating the efficient removal of redundant data. Additionally, ITAKE provides multiple processing options for dealing with specific types of unstructured text. In the realm of cleansing rule design, ITAKE employs a strategy that diversifies cleansing algorithms based on the distinct needs of various tasks and datasets.

3.2 LLM Fine-tuning and Extraction

Although LLMs have now developed rapidly and are widely used in knowledge extraction, they still perform poorly when oriented to specific domains, such as biomedical and financial domain, due to insufficient domain-specific training data(Keraghel et al., 2024). Therefore, we propose a method that integrates LLMs knowledge to enhance the performance of specific-domain models. Firstly, we improve the structure of the LLMs model to make it more adaptable to knowledge extraction and preserve the structural characteristics. Secondly, we

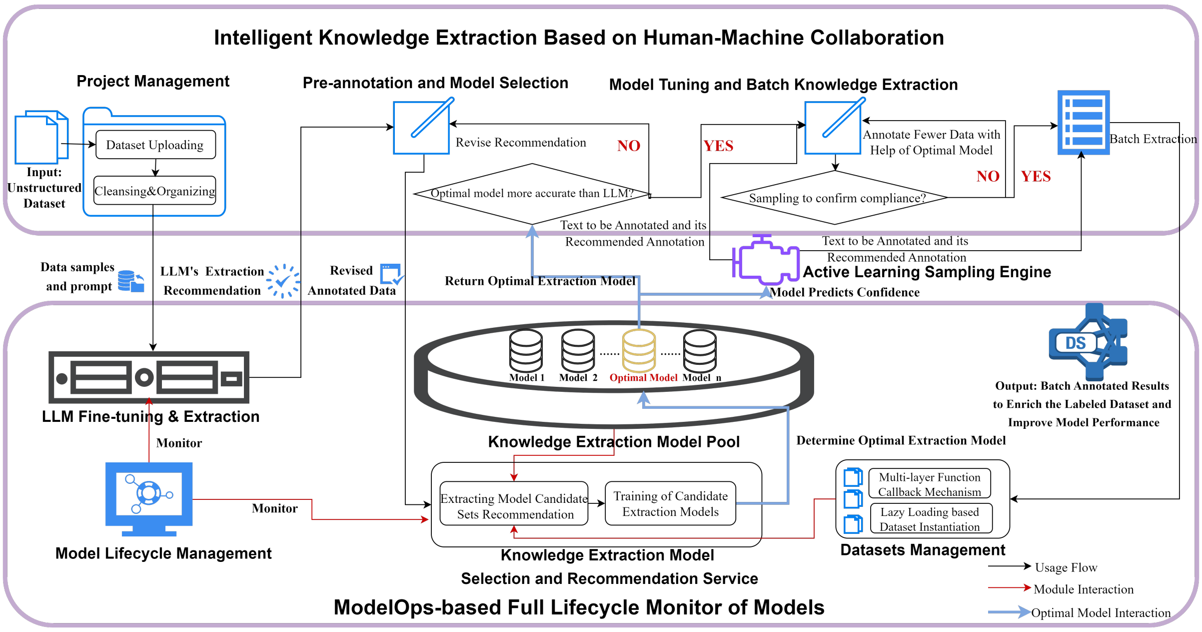

Figure 1: Architecture of ITAKE. Top: Intelligent Knowledge Extraction Based on Human-Machine Collaboration subsystem. Bottom: ModelOps-based Full Lifecycle Monitor of Models subsystem.

adopt the LoRA fine-tuning method and incorporate vocabulary information into the model training, making the training process more efficient. Finally, to fully utilize the fine-tuned LLM to enhance the specific-domain model, we convert the output of the LLMs into a knowledge concentration matrix and inject it into the model (Wang et al., 2023b). Specifically, after uploading the dataset, the user can select the LLM fine-tuned with data from the corresponding domain or similar domains according to the type of the uploaded dataset to be used as the base model for recommendation in the pre-annotation stage. It is important to note that during the subsequent knowledge extraction process, we will not fine-tune the LLMs using annotated data within the system. Instead, we will only utilize the LLMs API for inference. This approach is adopted because fine-tuning LLMs requires a substantial amount of annotated data and computational resources, which contradicts the objective of performing lightweight knowledge extraction tasks within ITAKE. Specifically, for LLMs already deployed on servers, we will employ a method similar to that of ChatGPT. The text requiring inference and the prompts will be transmitted to the LLMs via network requests using the LLM's native API in their deployment documents. This approach allows for the LLMs and ITAKE to be deployed on different servers, thereby reducing coupling and enhancing deployment efficiency and reusability.

3.3 Pre-annotation and Model Selection

To tackle the challenge of a scarcity of labeled data in specific fields, we employ LLMs for providing recommendations. In detail, upon the user engaging the "Get Large Language Model Recommendation" button, the extraction tool's backend transfers the present text along with its associated prompt to the LLM previously chosen, thereby acquiring a recommendation. Users are then tasked with revising these suggested outcomes. The modified results are subsequently forwarded to a candidate knowledge extraction model for its training. The criteria for selecting these alternative models will be elaborated upon in the subsequent section. The main annotation page is shown in **Fig.2**, which is similar to **3.5**.

3.4 Knowledge Extraction Model Recommendation and Selection Service.

This phase is divided into two stages: the recommendation of candidate models, and the selection of a model after the training of candidate models. Initially, to address the challenge of selecting appropriate knowledge extraction models, ITAKE has designed and implemented a dataset similarity-based model recommendation approach. This method employs Maximum Mean Discrepancy (MMD) (Gretton et al., 2006) and the Fréchet distance (FD) (Eiter and Mannila, 1994) to calculate similarities between datasets. These similarity metrics are then

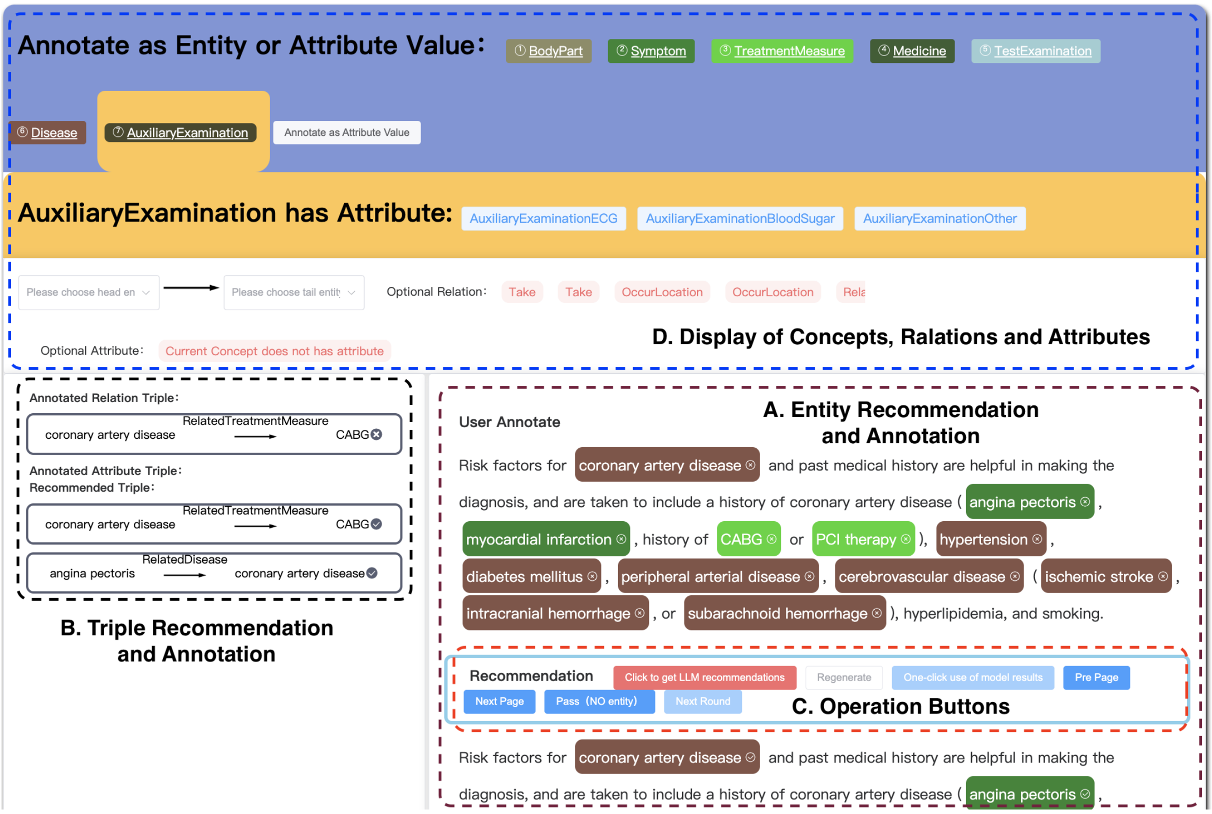

Figure 2: The main annotation page is divided into four main sections, which are **A.**Entity Recommendation and Annotation, **B.**Triple Recommendation and Annotation, **C.**Operation Buttons and **D.**Display of Concepts, Relations and Attributes. Users can manually annotate or use recommendations directly, which is detailed shown in video.

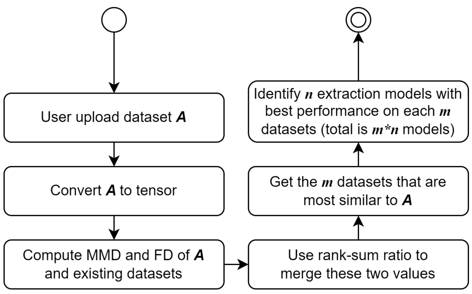

Figure 3: Workflow of Knowledge Extraction Model Recommendation

merged using the rank-sum ratio method to compute the overall dataset similarity.

Building on the computation of dataset similarity, the system devises a recommendation method for extraction models. It aims to recommend the optimal model for the uploaded dataset, thereby eliminating the need for repeated trials across numerous models, as illustrated in **Fig.3**. Specifically, for the uploaded dataset A, ITAKE identifies m datasets most similar to A through dataset similarity calculations. Subsequently, it identifies n extraction models with the best performance on each of these m datasets, where both m and n can be user-defined. After training the $m*n$ models with revised annotations, ITAKE ranks the candidate models based on various training metrics, such as precision and F1-score, facilitating user selection. The setting page is shown in **Fig.4**. Through this process, ITAKE provides users with more precise and targeted model recommendations, significantly reducing the time and effort users spend on model selection and adjustment.

3.5 Model Tuning and Batch Knowledge Extraction

When the accuracy of the optimal model surpasses LLM, the annotation process advances to the second phase: model tuning and batch knowledge extraction. At this stage, the model for knowledge extraction is the optimal model, selected by the user after comparing the training matrics of various candidate extraction models. The selection of unlabeled texts from ModelOps to be returned to the extraction subsystem is determined by an active learning sampling engine. Active learning is a research area within machine learning, employs sampling strategies to identify the samples

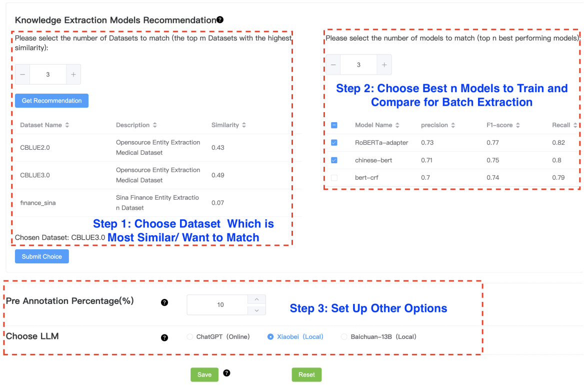

Figure 4: Pre-annotation settings can be set up in 3 steps as shown in the figure.

most beneficial for current model training (Shen et al., 2017; Settles, 2009). This approach aims to maximize model performance gains with a minimal number of samples, thereby reducing the data volume required to reach a predetermined performance benchmark.

To significantly reduce the total volume of text users must manually extract, ITAKE employs active learning methodology. We designs and tests various active learning sample selection strategies, encompassing strategies based on uncertainty, sample diversity, and a combination of both. Uncertainty-based strategies include the least confidence method (Agrawal et al., 2021), margin-based method (Balcan et al., 2007), and entropy-based method (Holub et al., 2008). The strategy based on sample diversity employs the K-means method (Vu et al., 2010), while the hybrid strategy integrates the gradient-based badge (Ash et al., 2019) method. The effect of active learning will be shown in Case Study and Evaluation. Once the model training meets the expected performance, ITAKE proceeds with the automatic batch extraction of the remaining texts, requiring users only to export the results without verifying.

Both parts **3.3** and **3.5** use models (LLMs or extraction models) for recommendation, which effectively reduces the user's labeling cost; at the same time, the system returns the higher quality extraction results annotated by the user to the model pool for model training, which ensures effective feedback from the human in the loop and enables the model accuracy to be steadily improved, thus solving **Challenge C2**. At the same time, these two parts integrate LLMs under low resources situation and use extraction models for well-labeled situations, ensuring a balance between efficiency and accuracy, thus addressing **Challenge C4**.

3.6 Dataset Management

Dataset management encompasses three key components: design of dataset specifications, implementation of multi-layer callback functions, and dataset instantiation via a lazy loading strategy. It is well known that, data standards serve as normative constraints that ensure uniformity, precision, and integrity of data, facilitating a common understanding, utilization, and exchange across various business systems. To streamline the integration for dataset providers and model developers, ITAKE adopts a unified dataset specification standard. It is important to underscore that ITAKE does not mandate users to pre-process the dataset to conform to this standard. Instead, it leverages a multi-layer callback function architecture to effectuate this transformation process.

Callback functions are a functional programming technique that encapsulates the logic of dataset processing and feature extraction into separate functions that are passed as arguments to other functions. This design allows the tool to dynamically change the processing flow at runtime for efficient adaptation between datasets and models. A common machine learning workflow in the dataset processing and model development phase is: acquiring data, data normalization, feature extraction, constructing a dataset class and a data loader. Based on this flow, ITAKE is designed with multiple layers of callback functions. In addition, in order to process data only when it is really needed (e.g., for model training, evaluation, or prediction), ITAKE employs a dataset instantiation method based on a lazy loading strategy.

3.7 Model Lifecycle Management

Users can monitor the performance of the model in real time, such as precision and F1-score. At the same time, they can track and monitor the training of the model in real time, such as CPU occupancy, memory information, etc. In addition, by combining with the dataset management module, the system can match and recommend the trained model based on the dataset similarity to be used for the recommendation of the results of the knowledge extraction, which greatly improves the re-usability of the model and the dataset. Through **3.4**, **3.6** and **3.7**, ITAKE provides effective reuse of models and datasets while providing management of full model lifecycle, thus addressing **Challenge C3**.

	Scope of Application		Technical				Model Service		Reusability	
Tools	**[A1]**	**[A2]**	**[B1]**	**[B2]**	**[B3]**	**[B4]**	**[C1]**	**[C2]**	**[D1]**	**[D2]**
Doccano	✓	✓	-	-	-	✓	-	-	-	-
MedCAT	-	-	-	✓	-	✓	-	-	✓	✓
FAMIE	✓	✓	-	✓	✓	✓	-	-	✓	✓
DeepKE	✓	✓	✓	✓	-	-	-	-	✓	✓
CollabKG	✓	✓	✓	-	-	-	-	-	-	-
Autodive	✓	✓	-	✓	-	✓	-	-	-	-
ITAKE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of some of the current knowledge extraction tools, selected on the basis of being popular or published in relevant conferences (e.g. ACL, EMNLP, etc.)

4 Evaluation and Case Study

4.1 Evaluation by Comparison with Other Tools

We compared ITAKE with some popular or already published annotation tools at relevant conferences, including Doccano (Nakayama et al., 2018), MedCAT (Kraljevic et al., 2021), FAMIE (Nguyen et al., 2022), DeepKE (Zhang et al., 2022b), CollabKG (Wei et al., 2023), Autodive (Du et al., 2023), to evaluate the system's performance. The comparison metrics discarded some traditional and commonly implemented features and instead focused on some innovative metrics as bellows: The first is **[A]. Scope of Application**, which includes *[A1]. Multidisciplinary* and *[A2]. Multilingual*. The second is **[B]. Technical**, which includes *[B1]. LLM*, *[B2]. Knowledge Extraction Model*, *[B3]. Active Learning* and *[B4]. Human-in-the-loop*. The third is **[C]. Model Service**, which includes *[C1]. Recommendation for What Model to Use* and *[C2]. Monitoring of Model*. The fourth is **[D]. Reusability**, which includes *[D1]. Reusability of Model* and *[D2]. Reusability of Dataset*. The comparison **Table 1** is as follows.

As can be seen from the comparison in the table, ITAKE's ability in model management and service is significantly better than other tools. In addition, ITAKE organically combines LLMs, extraction models, human-in-the-loop and active learning, which can significantly reduce costs and increase efficiency. Finally, ITAKE improves the reusability of datasets and models through dataset and model recommendation.

4.2 Case Study in Medical Knowledge Extraction

Knowledge extraction tasks play a crucial role in the healthcare domain by facilitating information structuring, feature extraction, and reasoning (Rajabi and Kafaie, 2022). Therefore, we carried out a batch of medical data knowledge extraction by cooperating with doctors from authoritative hospitals. Firstly, through the **Project Management** page, we uploaded the medical emergency guidelines to be annotated, while the ontology model was defined by professional doctors. After uploading the dataset, the **Dataset Management** module has already started the processing of the data in the background. The third step is to select our autonomously fine-tuned medical LLM called **Xiaobei**, which is fine-tuned by using medical knowledge on baichuan2-13b-chat (Baichuan, 2023) through **LLM Fine-tuning and Extraction**. In the fourth step **Knowledge Extraction Model Recommendation and Selection Service** module, the setting of m is 2, n is 3, and the recommended datasets are CBLUE2.0, CBLUE3.0 (Zhang et al., 2022a) where CBLUE3.0 is selected cause it has higher similarity. The three models corresponding to CBLUE3.0 are RoBERTa-adapter (Poth et al., 2021), BERT-CRF (Souza et al., 2019) and Chinese-BERT (Cui et al., 2020). After selecting the LLM and the extraction model to be used, we came to the fifth step of **Pre-annotation and Model Selection**. With a small amount of guidance from professional doctors, we asked 10 postgraduate medical students to annotate 400 texts with the entities recommended by Xiaobei, and trained all three models, with a training time of about $\mathbf{2.3}h$. The recall rates of the training were $\mathbf{73.7}\%$, $\mathbf{75.6}\%$, and $\mathbf{75.4}\%$, respectively, and thus BERT-CRF was finally selected as the final extraction model. In the sixth step of the **Model Tuning and Batch Knowledge Extraction** , we again asked students to annotate about 200 texts to train the BERT-CRF model. At this point, we sampled

Datasets	Random	Entropy	Least Confidence	Margin	Kmeans	Badge
CMeEE	10.14	**7.25**	17.39	14.49	8.70	11.59
CMeIE	42.25	**33.80**	47.89	50.70	42.25	46.62

Table 2: The percentage(%) of samples that need to be trained to reach the training target using different active learning approach. It can be seen that active learning can reduce the training data obviously while basically guaranteeing performance, while the Entropy-based sampling strategy uses the least amount of training data.

50 texts with model-recommended entities for expert checking, stopped manual confirmation after the recall rate reached $\mathbf{85}\%$, and directly performed batch automatic extraction on all remaining texts. In the end, we sliced **3,857** texts from 8 emergency guidelines and obtained **7,018** entity records from nine concepts: disease, clinical presentation, medical procedure, medical device, drug, medical test item, body, department, and microbiological class.

4.3 Evaluation of Active Learning

In order to reflect the effect of active learning in reducing data required for training, we first train the model using full data. On the CMeEE (Zhang et al., 2022a) dataset, the model achieves an optimal F1-score of $\mathbf{64.77}\%$ on the validation set, and on the CMeIE (Zhang et al., 2022a) dataset, the model achieves an optimal F1-score of $\mathbf{75.33}\%$ on the validation set for entity prediction, and $\mathbf{59.32}\%$ for relation prediction.

We then selected $\mathbf{90}\%$ of the performance of the model trained using the full amount of data as the targets and examined the percentage of samples that need to be trained to reach the training target using the active learning approach. The lower the percentage of samples needed, the more effective this active learning sampling strategy is. The experimental results are shown in **Table 2**.

5 Conclusion and Future Work

We developed ITAKE, a knowledge extraction system that combines LLMs and ModelOps. Its usability and cost reduction have been fully demonstrated through real case study. In the future, we hope to add events and multi-modal extraction, and add the LLMs self-feedback mechanism, so as to reduce human cost more effectively.

Limitations

As a knowledge extraction system, ITAKE lacks of support for nested, overlapping, or hierarchical entities, which is a complex and important aspect of the NER field. Besides, ITAKE does not facilitate collaborative use, limiting its applicability in complex and team-based settings.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U23A20468).

References

Ankit Agrawal, Sarsij Tripathi, and Manu Vardhan. 2021. Active learning approach using a modified least confidence sampling strategy for named entity recognition. *Progress in Artificial Intelligence*, 10:113–128.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. 2007. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Yi Du, Ludi Wang, Mengyi Huang, Dongze Song, Wenjuan Cui, and Yuanchun Zhou. 2023. Autodive: An integrated onsite scientific literature annotation tool. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 76–85, Toronto, Canada. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Thomas Eiter and Heikki Mannila. 1994. Computing discrete fréchet distance.

Óscar Fontenla-Romero, Bertha Guijarro-Berdiñas, David Martinez-Rego, Beatriz Pérez-Sánchez, and Diego Peteiro-Barral. 2013. Online machine learning. In *Efficiency and Scalability Methods for Computational Intellect*, pages 27–54. IGI global.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Alex Holub, Pietro Perona, and Michael C Burl. 2008. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE.

Waldemar Hummer, Vinod Muthusamy, Thomas Rausch, Parijat Dube, Kaoutar El Maghraoui, Anupama Murthi, and Punleuk Oum. 2019. Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*, pages 113–120. IEEE.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. Dreeam: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL, page (to appear), Dubrovnik, Croatia. Association for Computational Linguistics.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Minh Van Nguyen, Nghia Ngo, Bonan Min, and Thien Nguyen. 2022. FAMIE: A fast active learning framework for multilingual information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 131–139, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Clifton Poth, Jonas Pfeiffer, Andreas R"uckl'e, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Enayat Rajabi and Somayeh Kafaie. 2022. Knowledge graphs and explainable ai in healthcare. *Information*, 13(10):459.

Burr Settles. 2009. Active learning literature survey.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Viet-Vu Vu, Nicolas Labroche, and Bernadette Bouchon-Meunier. 2010. Active learning for semi-supervised k-means clustering. In *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, volume 1, pages 12–15. IEEE.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Zhiyuan Wang, Qiang Zhou, Zhao Junfeng, Yasha Wang, Hongxin Ding, and Jiahe Song. 2023b. A knowledge-enhanced medical named entity recognition method that integrates pre-trained language models. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pages 296–301. IEEE.

Xiang Wei, Yufeng Chen, Ning Cheng, Xingyu Cui, Jinan Xu, and Wenjuan Han. 2023. Collabkg: A learnable human-machine-cooperative information extraction toolkit for (event) knowledge graph construction. *arXiv preprint arXiv:2307.00769*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022a. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Ningyu Zhang, Xin Xu, Liankuan Tao, Haiyang Yu, Hongbin Ye, Shuofei Qiao, Xin Xie, Xiang Chen, Zhoubo Li, and Lei Li. 2022b. Deepke: A deep learning based knowledge extraction toolkit for knowledge base population. In *EMNLP (Demos)*, pages 98–108. Association for Computational Linguistics.