# Active Learning for Indexing Research Objects in Scientific Literature: A Postprint Study

**Authors:** He Huixin, Liu Lijuan

**Date:** 2017-10-11T00:00:00+00:00

## Abstract

Objective: To identify instances of research object attributes in paper titles, with the aim of maximizing the accuracy of research object recognition using a minimal number of annotated samples. Method: This study analyzes the grammatical features of research objects in scientific literature, employs a small sample set to identify and extract research objects based on the Conditional Random Fields (CRF) sequence labeling algorithm, and introduces an iterative annotation framework based on active learning with unlabeled data to enhance the accuracy of research object recognition. Results: The proposed approach can efficiently leverage unlabeled data and maximize recognition accuracy, achieving an annotation accuracy of 78.3%. Limitations: The computational efficiency of the algorithm requires further optimization. Conclusion: The method exhibits favorable recognition performance for instances of research object attributes in scientific literature, establishing a foundation for further excavation of the knowledge system and structure within scientific literature.

## Full Text

## Preamble

ChinaXiv Collaborative Journal, Issue 268, 2016, No. 3

**An Active Learning-Based Labeling System for Research Objects in Scientific Literature**

He Huixin, Liu Lijuan
(Tongfang Knowledge Network Technology Co., Ltd. (Beijing), Beijing 100192, China)

## Abstract

[**Objective**] This study aims to identify research object attribute instances from paper titles, seeking to maximize the accuracy of research object recognition us-

ing a minimal number of labeled samples. [**Methods**] We first analyzed the grammatical characteristics of research objects in scientific literature. Using a small sample set, we applied the Conditional Random Fields (CRF) sequence labeling algorithm to recognize and extract research objects. We then introduced an active learning-based iterative labeling framework leveraging unlabeled data to improve recognition accuracy. [**Results**] The proposed method efficiently utilizes unlabeled data and maximizes research object recognition accuracy, achieving 78.3% labeling accuracy. [**Limitations**] The algorithm' s computational efficiency requires further optimization. [**Conclusions**] The method demonstrates effective identification of research object attribute instances in scientific literature, laying a foundation for automated extraction and organization of knowledge structures from academic papers.

**Keywords:** Scientific literature; Research objects; Conditional Random Fields; Iterative labeling system; Active learning
**Classification Codes:** TP393; G25

## 1. Introduction

Scientific papers represent textual manifestations of new research findings or innovative insights on academic topics, serving as a formal medium for researchers to summarize and present their work. A typical scientific paper comprises various research elements, including background, research objects, processes, methods, and conclusions. The research object refers to the core subject under investigation—the primary focus that efficiently and clearly positions the paper' s scope, encompassing entities such as objective things, theories, events, processes, and relationships. Extracting research objects presents the main research target in an intuitive form, enabling readers to quickly grasp relevant information and facilitating retrieval and comparison of related studies.

This paper addresses the identification and extraction of research object attribute instances from academic papers, proposing an active learning-based labeling framework. By using a small set of labeled samples for research object annotation while fully leveraging large amounts of unlabeled data, our approach minimizes manual annotation costs while maximizing extraction accuracy. This system provides a valuable reference for automated extraction and organizational management of knowledge structures in academic papers.

Research object extraction falls under the broader scope of attribute extraction, which itself belongs to fine-grained knowledge extraction. Regarding extraction target types, attribute extraction primarily focuses on two categories: entity attribute extraction (e.g., people [1-2], products [3-4]) and concept attribute extraction [5-6]. Concept attribute extraction can be further divided into general concept attribute extraction [7] and academic concept attribute extraction [5-8]. This study targets medical domain papers, where extraction objects include both academic concept attribute instances and medical named entities [9] such as diseases, drugs, and treatment methods. Therefore, we approach the problem

from two perspectives: academic concept attribute extraction and named entity recognition.

Methods for attribute extraction in domain-specific literature primarily include rule-based approaches, machine learning methods, and hybrid approaches combining both.

### 2.1 Rule-Based Methods

Rule-based methods employ manually or automatically constructed rules to identify linguistic patterns between relations and concepts, formulating extraction rules accordingly. Fundel et al. [10] developed candidate relation rules to extract gene-protein relationships from Medline abstracts, using rule-based methods with filtering mechanisms to achieve attribute extraction. However, since candidate relations were manually selected, the approach has inherent limitations, and accuracy is also affected by the performance of existing word segmentation tools. Zhang et al. [11] utilized association rules for knowledge extraction from medical data, analyzing word co-occurrence patterns to extract semantic relationship patterns between main and sub-topic words for four types of anti-tumor drugs, thereby achieving attribute relation extraction.

### 2.2 Machine Learning Methods

Machine learning approaches transform attribute extraction into classification or sequence labeling problems, requiring specific features and pre-labeled training data. Conditional Random Fields (CRFs) are probabilistic graphical models [12] that effectively capture long-distance dependencies between elements while avoiding the label bias problem present in Maximum Entropy Markov Models (MEMM) and other conditional Markov models. Meng et al. [13] applied CRF-based term recognition to the *Shanghan Lun* medical text, conducting comparative experiments on features to build an automatic Chinese medical term recognition model. Zhang et al. [14] identified attribute instances corresponding to topics in innovation sentences by leveraging hierarchical topic words recognized through domain ontologies or lexicons, combining semantic annotation, dependency parsing, and domain ontology attribute classes to improve recognition accuracy.

### 2.3 Hybrid Rule and Machine Learning Methods

Attribute extraction for concepts or named entities has wide applications across disciplines. In natural sciences such as computer science [15], natural sciences [16], and molecular materials science [17], attribute extraction plays important roles in ontology construction, question-answering systems, and automatic summarization. Pham et al. [15] used a ripple-down rule-based approach to establish text annotation rules, applying regular expression-based filtering and analyzing specific features of different annotation categories to determine classification results. Pechsiri et al. [16-17] focused on causal attribute relationships, analyzing

verb connectors in causal relations to annotate verbs, causes, and effects, and used a Bayesian classifier to determine whether verb-linked descriptions represented causes or effects. Xiao et al. [18] studied the environmental impact of nanomaterials, pre-selecting six entity types and three attribute types related to nanotoxicity, extracting entity-attribute relationships and attribute values at the paragraph level.

In social sciences, some academic concept attributes are abstract or subjective [8], requiring identification of various features such as linguistic description characteristics and positional distribution features before extraction. Ding et al. [8] manually constructed nine major categories of rules for attribute extraction and applied them to academic concept extraction from papers published in *Journal of the China Society for Scientific and Technical Information*. Cheng et al. [19] utilized Bootstrapping methods and pattern-based named entity recognition for semi-supervised named entity recognition in specific domains.

Regarding training processes in machine learning methods, they can be categorized into supervised learning [20-21], semi-supervised learning [22-23], and unsupervised learning. Supervised learning relies entirely on manually annotated training data for parameter estimation, requiring large labeled datasets to ensure generalization capability, which is extremely time-consuming and labor-intensive. Particularly for domain-specific academic annotation, annotators typically require background knowledge. Unsupervised learning requires no manual annotation but often yields poor results due to rule limitations. Semi-supervised methods combine both approaches, leveraging their respective advantages to reduce manual annotation costs while improving attribute extraction accuracy.

Drawing upon these attribute extraction achievements and referencing machine learning research, particularly on active learning and CRF sequence labeling, this study explores how to use a small number of labeled samples for research object attribute annotation. We investigate how to estimate thresholds for selecting valuable samples from large unlabeled collections for manual annotation, minimizing labor costs while achieving optimal performance. By constructing character-based features for CRFs to train models for research object extraction and employing active learning for threshold estimation, we iteratively label datasets to improve accuracy.

## 3. Research Object Generation and Labeling Framework

Paper titles provide the most concise and logical combination reflecting the most important research content, including keywords that profoundly reveal the main research content and provide specific practical information for retrieval. Therefore, paper titles serve as the primary target for research object extraction.

We extract research objects from representative conceptual attributes in paper titles by analyzing their semantic and positional characteristics. Through semantic annotation of titles and an active learning-based labeling generation framework, we use a small set of labeled samples for research object attribute annota-

tion while fully leveraging large amounts of unlabeled data to maximize labeling accuracy. We compare our approach with Hidden Markov Model (HMM)-based extraction methods.

### 3.1 Rule-Based Research Object Extraction Strategy

For Chinese medical academic papers, we analyze paper titles and research objects, extracting subsequence strings consisting of single or multiple consecutive characters from titles as research objects. The rule-based strategy involves extracting research objects using manually crafted rules. Annotation results show that most research objects are medical terms or relationships extracted from titles, segmented by removing common words and using conjunctions and prepositions. We therefore construct a common word list to filter out such words from titles and segment titles using function words (prepositions, conjunctions, particles) to extract one or more research objects. Additionally, we write regular expressions to filter titles with obvious syntactic structures that were incorrectly labeled by the model.

This approach achieved certain effectiveness, correctly segmenting some data using conjunctions, prepositions, and particles. However, its limitations are shown in :

** Limitations of Rule-Based Strategy**

1. Some words function as common words in some titles but not in others, limiting dictionary matching effectiveness.
2. Segmentation using prepositions, conjunctions, and particles without considering semantic context fails when two connectives simultaneously modify one word.
3. Due to diverse syntactic patterns, simple word segmentation and rule-based methods cannot achieve ideal results.

### 3.2 CRF-Based Sequence Labeling Algorithm

Conditional Random Fields (CRFs), proposed by Lafferty et al. [12] in 2001, are probabilistic structural models for labeling and segmenting sequential data that have been widely applied in natural language processing.

Our basic labeling algorithm employs CRF-based sequence labeling [12], where feature selection most significantly impacts accuracy. The fundamental feature unit is the "character." We added features for each character as shown in :

** Features for CRF-Based Sequence Labeling Algorithm**

- **Character type**: Includes Chinese characters, digits, letters, punctuation, and uppercase numbers, plus the character' s relative position in the title (numeric value between 0 and 1).
- **Part-of-speech and position**: After segmenting the title, the part-of-speech of the word containing the character, and the character' s position

within that word (beginning, middle, or end).
- **High-frequency dictionary words**: Whether words from a high-frequency dictionary built from training corpus appear in the character' s sentence, and their distance from the character (e.g., words like "discuss," "based on" ).
- **Last 4 characters**: The final four characters of the title containing the current character.
- **Unigram**: With the current character at position 0, includes character groups at positions (-2, -1, 0, 1, 2).
- **Bigram**: With the current character at position 0, includes character pair combinations at positions (-2-*1*, -*1*0, 0*1*, *1*2).

### 3.3 Active Learning-Based Research Object Labeling Framework

The active learning process works as follows: Given a labeled dataset $K$ (possibly empty initially) and an unlabeled dataset $U$, we use information from $K$ to select a subset $C$ from $U$, request expert annotation for $C$, and add it to $K$ for the next iteration.

We employ an active learning approach where a CRF model trained on labeled data predicts labels for unlabeled data. We select minimal samples from the large unlabeled set for manual annotation, add them to the training set, and retrain the model iteratively to maximize accuracy.

During labeling with each trained CRF model, the classification stage compares probabilities of each character belonging to different classes, selecting the class with maximum probability. The difference between the highest and second-highest probabilities serves as a confidence measure for model classification. We select new data for labeling based on analyzing this probability difference.

#### (1) Probability Difference Analysis

To calculate the optimal threshold, we divide data into three groups: training data $a$ for model training, additional data $b$ for filtering and adding to the training set, and test data $c$ for comparing accuracy differences. Training on $a$ and predicting on $c$ yields accuracy $P_1$. We then select data from $b$ where $\varepsilon \in \partial$, add them to $a$, retrain, and predict on $c$ to obtain accuracy $P_2$. The difference $\Delta P = P_2 - P_1$ indicates whether data in this interval benefits the model. To minimize manual annotation costs, we must reduce the number of samples $N$ selected for labeling. We introduce the following evaluation formula:

$$R = \arg\max_{\varepsilon} \frac{\Delta P}{N} = g(\varepsilon)$$

where $\varepsilon$ = maximum probability − second maximum probability, $\Delta P$ represents the accuracy difference between two models, and $N$ is the number of manually labeled samples added. Larger $\Delta P$ and smaller $N$ yield better results with minimal data addition. The optimal $\varepsilon$ occurs when $R$ is maximized.

**(2) Active Learning Iterative Training Process**

The iterative training process filters unlabeled data using the threshold-estimated $\varepsilon$ value, manually annotates selected samples, and adds them to the training set for parameter re-estimation. Through multiple iterations, we balance data volume and accuracy, maximizing $R$. The workflow of the active learning-based research object labeling framework is shown in [Figure 1: see original paper].

**[Figure 1: see original paper] Workflow of Active Learning-Based Research Object Labeling Framework**

The process uses initial training data for model parameter estimation and evaluates accuracy on test data. Through active learning iterations, the established model filters unknown data, selects samples using threshold-estimated intervals, adds them to the training set, and re-estimates parameters. After multiple iterations, the model achieves optimal accuracy and training efficiency.

## 4. Experiments

### 4.1 Data Source

Experimental data were sourced from medical academic papers in CNKI [24]. We manually annotated research objects in paper titles, selecting 18,449 entries as initial training data.

### 4.2 Experimental Setup and Results Analysis

**(1) Threshold Estimation Experiment**

The accuracy differences between consecutive experiments are shown in [Figure 2: see original paper].

**[Figure 2: see original paper] Accuracy Difference Between Consecutive Experiments**

Training data changes are illustrated in [Figure 3: see original paper].

**[Figure 3: see original paper] Training Data Changes**

The trend of $R$ values is shown in [Figure 4: see original paper].

**[Figure 4: see original paper] Trend of $R$ Values**

Results demonstrate the relationship between accuracy difference and added data volume. As the threshold increases from 0 to 0.3, the number of added labeled samples grows and model accuracy rises, indicating that data in this interval significantly supplements the model. When the threshold is in [0.3, 0.8], accuracy growth slows. In [0.8, 1], accuracy fluctuates, suggesting data in this interval introduces noise.

Data volume increases linearly with the threshold. Calculations show that $\varepsilon \in [0, 0.3]$ improves accuracy by 1.7 points, while $\varepsilon \in [0.3, 0.8]$ improves it by only 1.3 points, with a data volume ratio of approximately 1:3. Since our goal is to minimize manual annotation costs while maximizing accuracy, data in $\varepsilon \in [0, 0.3]$ provides the greatest model improvement with the smallest labeling effort. The significant accuracy gain in this interval justifies its selection.

[Figure 4: see original paper] shows that $R$ increases when $\varepsilon \in [0, 0.3]$ and decreases when $\varepsilon > 0.3$, confirming that selecting the [0, 0.3] interval minimizes data volume while maximizing accuracy.

**(2) Iterative Training Experiment and Comparative Analysis**

In single experiments, we performed five-fold cross-validation on initial training data, training CRF models on each fold and evaluating on corresponding test sets, with final accuracy averaged across folds. We tested CRF window sizes of 2, 4, 6, 8, 10, and 12. Accuracy improved with larger windows, but when window size reached 8-12, performance gains became negligible while training time and memory usage increased exponentially. Therefore, we selected the optimal window size of 6. The CRF algorithm used L-BFGS parameter estimation with L1/L2 regularization coefficients $c_1 = 0$ and $c_2 = 1$.

The iterative training process filters unknown data using the threshold $\varepsilon \in [0, 0.3]$ from the above experiment, manually annotates selected samples, and adds them to the training set for parameter re-estimation. Through multiple iterations, we balance data volume and accuracy, maximizing $R$. Using our proposed active learning iterative labeling framework, accuracy trends are shown in [Figure 5: see original paper].

**[Figure 5: see original paper] Accuracy Change Trend**

On the manually annotated data, the initial model achieved 67.5% accuracy. Using pure CRF labeling yielded preliminary results. As iterations progressed, accuracy $P$ increased within a certain range, demonstrating that active learning adds targeted data to address model blind spots. As the feature space approaches completeness, the growth rate slows. After five iterations, research object extraction accuracy reached 78.3%, representing substantial improvement.

Comparison with HMM-based extraction is shown in [Figure 6: see original paper].

**[Figure 6: see original paper] Comparative Experimental Results**

The active learning CRF sequence labeling method outperformed HMM-based methods at every data segment, with overall performance significantly superior. As data volume increased, the active learning CRF method showed clear accuracy improvements. HMM' s assumptions are suitable for small datasets but inadequate for real-world corpora where observation sequences exhibit multiple interacting features. Due to the structural complexity of entities, simple feature functions cannot capture all characteristics, limiting HMM' s effectiveness.

Combined results from [Figure 5: see original paper] and [Figure 6: see original paper] demonstrate that our method significantly improves accuracy with increasing iterations, validating its effectiveness compared to HMM and pure CRF approaches.

## 5. Conclusion

This study systematically analyzed the structural and semantic characteristics of research objects in domain-specific scientific literature, using CRF sequence labeling to extract research objects from academic papers. We proposed an active learning-based research object labeling framework that selects informative samples from unknown datasets to optimize model performance. Our method reduces manual annotation costs, maximizes machine learning efficiency, fully leverages large amounts of unlabeled data, and achieves optimal performance for research object extraction. The approach is applicable not only to medical literature but also to metadata extraction in other domains, providing strong guidance for other metadata extraction tasks.

## References

[1] Lan M, Zhang Y Z, Lu Y, et al. Which Who are They? People Attribute Extraction and Disambiguation in Web Search Results [C]. In: Proceedings of the 18th World Wide Web Conference, Madrid, Spain. 2009.

[2] Li Hongliang. Research on Character Attributes Extraction Based on Rules from Baidu Encyclopedia [D]. Chengdu: Southwest Jiaotong University, 2013.

[3] Zeng Daojian, Lai Siwei, Zhang Yuanzhe, et al. Open Entity Attribute-Value Extraction from Unstructured Text [J]. Journal of Jiangxi Normal University: Natural Science Edition, 2013, 37(3): 279-283.

[4] Ghani R, Probst K, Liu Y, et al. Text Mining for Product Attribute Extraction [J]. ACM SIGKDD Explorations Newsletter, 2006, 8(1): 41-48.

[5] Jia Zhen, Yang Yufei, He Dake, et al. Attribute and Attribute Value Extracted from Chinese Online Encyclopedia [J]. Acta Scientiarum Naturalium University Pekinensis, 2014, 50(1): 41-47.

[6] Liu Lijia, Guo Jianyi, Zhou Lanjiang, et al. Domain Concepts Entity Attribute Relation Extraction Based on LM Algorithm [J]. Journal of Chinese Information Processing, 2014, 28(6): 216-222.

[7] Ding Yufei, Wang Yuefen, Liu Weijiang. Research on Knowledge Extraction for Semi-structure Text [J]. Information Studies: Theory & Application, 2015, 38(3): 101-106.

[8] Ding Junjun, Zheng Yanning, Hua Bolin. Academic Concept Attribute Extraction Based on the Rules [J]. Information Studies: Theory & Application, 2011, 34(12): 10-14.

[9] Rebholz-Schuhmann D. Biomedical Named Entity Recognition, Whatizit [A]. // Encyclopedia of Systems Biology [M]. Springer New York, 2013: 132-134.

[10] Fundel K, Küffner R, Zimmer R. RelEx—Relation Extraction Using Dependency Parse Trees [J]. Bioinformatics, 2007, 23(3): 365-371.

[11] Zhang Han, Lu Zhenyu, Cui Lei. Knowledge Extraction from Medical Literature Database Using Association Rule Mining —Taking Four Anti-neoplastic Medicines as an Example [J]. New Technology of Library and Information Service, 2006(9): 49-52.

[12] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. In: Proceedings of the 18th International Conference on Machine Learning. 2001.

[13] Meng Hongyu, Xie Qingyu, Chang Hong, et al. Automatic Identification of TCM Terminology in Shanghan Lun Based on Conditional Random Field [J]. Journal of Beijing University of Chinese Medicine, 2015, 38(9): 587-590.

[14] Zhang Fan, Le Xiaoqiu. Research on Recognition of Concept Attribute Instances in Innovation Sentences of Scientific Research Paper [J]. New Technology of Library and Information Service, 2015(5): 15-23.

[15] Pham S B, Hoffmann A. Extracting Positive Attributions from Scientific Papers[A]. // Discovery Science [M]. Springer Berlin Heidelberg, 2004: 169-182.

[16] Pechsiri C, Kawtrakul A. Mining Causality from Paragraphs for Question Answering System [A]. // Proceedings of the 2008 International Conference on Computer and Electrical Engineering. 2008: 213-217.

[17] Pechsiri C, Kawtrakul A. Mining Causality from Texts for Question Answering [A]. // Proceedings of the 2007 International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. 2007: 935-938.

[18] Xiao J, Su J, Zhou G D, et al. Protein-Protein Interaction Extraction: A Supervised Learning Approach [C]. In: Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine. 2005: 51-59.

[19] Cheng Ziguang. Research on Named Entity Recognition and Relation Extraction Facing to Domain-Oriented Knowledge Base Construction [D]. Harbin: Harbin Institute of Technology, 2014.

[20] Zhang Yijia. Information Extraction in Biomedical Literature and Protein Complex Identification [D]. Dalian: Dalian University of Technology, 2014.

[21] Li Y P, Hu X H, Lin H F, et al. Learning an Enriched Representation from Unlabeled Data for Protein-Protein Interaction Extraction [J]. BMC Bioinformatics, 2010, 11(S2): 7-10.

[22] Yan Zifei, Ji Donghong. Exploration of Chinese Temporal Information Extraction Based on CRF and Semi-supervised Learning [J]. Computer Engi-

neering and Design, 2015, 36(6): 1642-1646.

[23] Xiao C, Su J, Zhou G D, et al. Protein-Protein Interaction Extraction: A Supervised Learning Approach [C]. In: Proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China. 2013: 25-30.

[24] CNKI [OL]. [2015-06-25]. http://www.cnki.net/.

**Author Contributions**

He Huixin: Conceived the research idea, designed the study, revised the final manuscript.
Liu Lijuan: Collected, cleaned, analyzed, and processed data; conducted experiments; drafted the manuscript.

**Conflict of Interest Statement**

All authors declare no conflict of interest.

**Supporting Data**

Supporting data [1-3] are available in the online version of the journal at http://www.infotech.ac.cn; supporting data [4] is self-archived by the authors, E-mail: huixinhe@qq.com.

[1] He Huixin, Liu Lijuan. tools_url.txt. Web links for research tools crfsuite and python-crfsuite.
[2] He Huixin, Liu Lijuan. title.xlsx. Original experimental paper title data.
[3] He Huixin, Liu Lijuan. X.txt. Feature-converted experimental paper title data.
[4] He Huixin. studyObject.xlsx. Manually annotated research object data from experimental paper titles.

**Received:** October 13, 2015
**Revised:** December 22, 2015

*Note: Figure translations are in progress. See original paper for figures.*