# A Unsupervised Method for Terminology Extraction from Scientific Text

Wei Shao
1600016634@pku.edu.cn
Department of Information Management, Peking University
Beijing, China

Bolin Hua
huabolin@pku.edu.cn
Department of Information Management, Peking University
Beijing, China

## Abstract

A lot of new scientific documents are published on various platforms everyday. It is more and more important to quickly and efficiently find new terminologies from these documents. However, most related works rely on labeled data and hard to deal with unlabeled new documents well. For this, we introduce a unsupervised method based on sentence patterns and part of speech sequence. Our method just needs a few initial learnable patterns to obtain initial terminology tokens and their part of speech sequence. In this process, new patterns are constructed and can match more sentences to find more part of speech sequences of terminology. Finally, we use the part of speech(POS) set and sentence patterns to match terminologies in new scientific text. Experiments on paper abstracts from Web of Knowledge show that this method are practical and achieve a good performance on our testing data.

***CCS Concepts:*** • **Information systems** → *Data mining*; **Information extraction**; • **Applied computing** → Document management and text processing.

***Keywords:*** terminology extraction, unsupervised method, scientific text

## 1 Introduction

With the quick development of science and technology, more and more papers are produced every day. So it is more and more difficult for researchers to find something new by reading papers. How to find new terminologies from new published papers automatically with computer becomes an important problem.

Generally, finding new terminology relies on named entity recognition(NER). However, many high performance methods(reference) need supporting of labelled data. Although they can obtain excellent results on training and testing data, it is hard for them to process new unlabelled data that we often face. One factor leading to this gap is that features of new scientific document text are different from features models learn on training data owing to the difference between their domains. Also, these new scientific texts are usually lack labels for extraction. So a unsupervised method which can also adapt different domains is need.

To overcome this difficulty, we propose a unsupervised method based on sentence pattern and part of speech. In detail, we initialize a few patterns to extract terminologies in certain sentences. In this step, we can obtain some terminologies and their part of speech sequences with some natural language processing tools (nltk[7], stanfordnlp[9], etc). Then, we try to find the same POS sequences in sentences not matched by initial patterns with obtained terminologies' POS sequences. If a sentence is matched, we will utilize suitable words in this sentence to replace the extendable parts of initial patterns. In this case, we obtain new patterns and are able to use these new patterns to match other sentences to get more terminologies. After several iterations, most terminology in scientific sentences can be extracted. Result shows that we can get a high performance on unlabeled texts from paper abstracts from Web of Knowledge.

In summary, we propose a unsupervised method for terminology extraction from scientific texts, which partly solve the difficult of extracting from unlabelled and different domains' data. Experiments show that our method can achieve a high accuracy(0.64) on unlabelled data from Web of Knowledge.

## 2 Related Work

Recent years, terminology extraction has attracted more and more attention. And all kinds of methods are produced to achieve a better performance.

Some methods rely on string, syntax and other original features. Liu li[5] and Zen Wen[12] use length of word and grammatical features to choose terminology candidates.

External resources such as dictionary[13], lexicon[1], parallel corpora[4], wikipedia[14] and so on are often used to extract terminologies from unlabelled data. However, this methods have a low performance when they deal with domains with low or no resources supports.

With the development of deep learning and machine learning, some methods based on machine and deep learning are put forward. Among these methods, LSTM[2] and CRF[10] and their variants achieve the best performance. Although these methods can obtain a better results, they rely on a large number of labelled data and have a poor performance on new unlabelled data.

To solve the gap between training data and new practical data, some semi-supervised and unsupervised methods

are proposed. A graph-based semi-supervised algorithm[8] working with a data selection scheme to leverage unannotated data achieve a high F1 on SemEval Task 10 ScienceIE task. Automatic rule learning based on morphological features method[11] also used to extract entities which doesn't need any annotated data. However, owing to the difficulty of searching optimal parameters, these methods can't get fully developed.

Besides single method or algorithm, terminology extracting systems utilizing all kinds of methods and are practical in real world are also focused by researchers. Xu Hao[3] put forward an extracting system scheme based on traditional processing. Yu Li[6] designed a system which uses seed terminology words from scientific database to create annotated data and train a deep learning extracting model with these labelled data. Both these systems can achieve high accuracy or recall.

In view of the fact that high performance methods hard to deal with unlabelled text and many unsupervised methods rely on external resources, we propose a unsupervised method based on POS sequences and sentence pattern to extract terminology entities from scientific text. This work is also beginning of a unsupervised extraction system.

# 3 Method

## 3.1 Overview

Our method aim to extract terminology from unlabelled scientific texts. To achieve this goal, implicit features of sentence are taken into full consideration. In detail, we utilize two features of terminology. One is the surrounding words and another is POS sequences of terminology.

The process can be divided into two steps. One step is to cool start our model with unlabelled data. In this step, the model will learn sentence patterns, POS sequences of terminology from data. Besides the sentences text, the POS of each token in the sentences is also needed in learning process. Another step is to extract terminology with learned model. For a new sentence, the model can extract terminology with sentence pattern when only sentence string is input. Also, the model can use POS sequences to extract terminology entities if the sentence's POS sequence is also input with sentence string.

**3.1.1 First Step.** First, we use tools to obtain POS sequences of sentences. Then a few patterns are used to match sentences. The patterns are specially designed regular expressions consisting of special words and matching groups. When the pattern match a sentence, we can get terminology string from matched groups. After filtering and post-process, suitable terminology tokens and their POS sequences are output.

Second, POS sequences of extracted terminology tokens are used to match sentences which are not matched. Once sentence POS sequences contains terminology POS sequences, the sentence is matched. Then we use to certain tokens in sentences to replace the extendable part in patterns to produce new patterns. After that, new found patterns will try to match not matched sentences. After several iterations, we can get new patterns and terminologies and their POS sequences.

**3.1.2 Second Step.** Finally, for new sentences, we can use sentence patterns obtained from previous iterations to extract terminology from them. Also, we can utilize the POS sequences to find terminology with sentence's POS sequence input.

## 3.2 Pattern Description

The sentence pattern used to extract target entities from sentence string is a kind of regular expression.

```
Pattern One
r"(.+?) (?:is|was|are|were) proposed (?:by|to|for|with|that)", 1, "proposed"

Pattern Two
r"(?:we|to|and|then|here) (?:propose|proposed) (.+?) (?:by|to|for|with|that)"
```

**Figure 1.** Pattern Example

Examples are given in figure.1. These are two patterns aiming to extract method terminology. "propose" is a word which often appear with method words at the same time. Border words like "by, to, for" are used to limit the range of terminology words. What we want is matched by (.+?).

For generation of new patterns, we can use words from matched sentence to replace the extendable part of extant pattern. For pattern one and pattern two in figure.1, the extendable part is "propose" and "proposed". They can be replaced by "develop", "present", "put forward" and so on. In this case, new patterns are obtained and can be used to extract terminology in other sentences.

## 3.3 Cool Start

The process of cool starting of our method is shown in figure.2.

The input of the method are sentences including their POS sequences from scientific texts. First, we use each pattern from initial pattern base to match each sentence from sentence base. If the matching is successful, the sentence will be moved to extracted sentence base and we can obtain terminology words and their POS sequences. Otherwise, the sentence will be moved to unextracted sentence base. After getting terminology words and their POS sequences, we need to filter them to obtain more accurate results. The filtered POS sequences are moved to POS Sequence Base. Then, for each POS sequences of POS sequence base, it is used to find if the sentence POS sequence in unextracted sentence base contains itself. If sentence POS sequence contains, we can choose the candidate words from matched sentence for generation of new patterns. After new patterns are generated,

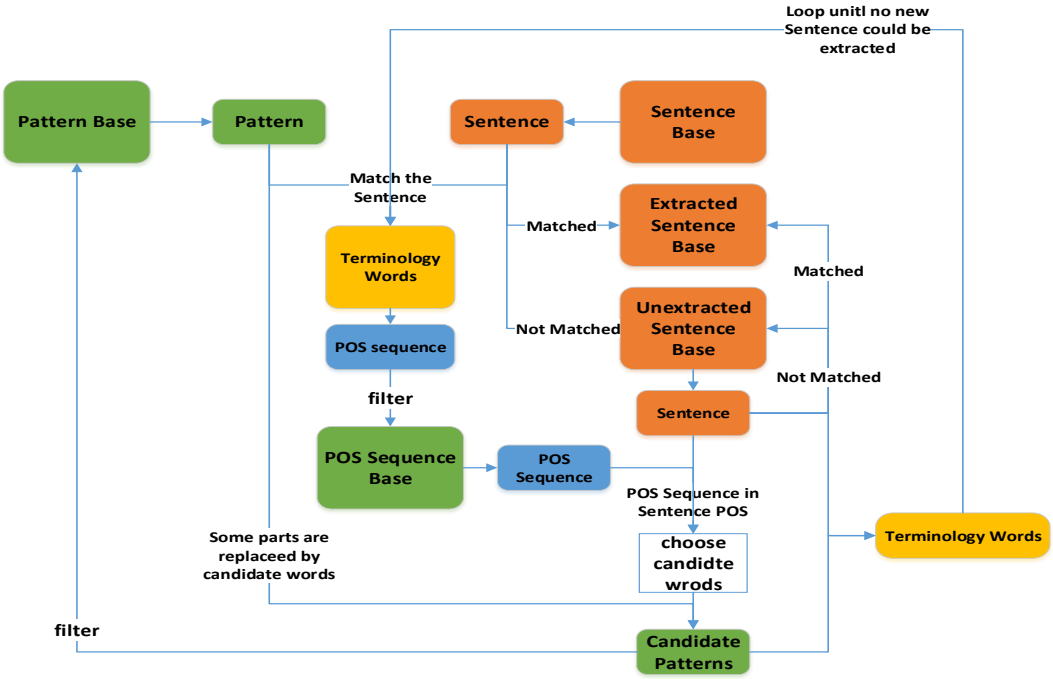


**Figure 2.** Cool Start Process

we use them to match sentences in unextracted sentence base and new terminology words are obtained. Then we can filter new generated patterns according to their matching results and move suitable patterns to pattern base. For new terminology words, they replace the initial extracted terminology words to participate in the extraction loop until no new sentence could be extracted.

### 3.4 Extraction for New Data

After the method learned on unlabelled data, we can obtain sentence patterns and POS sequences of terminology words. Here are two approaches to get new terminologies from new unlabelled data.

One is that we can use patterns to match sentences for obtaining new terminologies when only sentence string is input. Another is that when sentence string and POS sequence (processed by natural language tools) are input, we can use POS sequence to match POS sequence of sentences to get a more accurate result.

## 4 Experiment and Result

### 4.1 Data and Preprocessing

To evaluate the performance of our method, we crawled 200k+ scientific abstracts from Web of Knowledge. These abstracts are from different domains including machine learning, big data and data mining.

As for preprocessing, we utilize nltk to split abstracts into sentences and splitted sentences into tokens. Also we use stanfordnlp to get POS tags and dependency relations of cut sentences. Finally, data consists of four parts: sentences of abstracts, tokenized sentences of abstracts, POS tags of sentences, dependency relations of sentences. In this method, we use the tokenized sentences of abstracts and POS tags of sentences.

In experiment, we use 54000 sentences and their POS sequences as training data and 1000 sentences and their POS sequences as testing data. All sentences are unlabelled.

### 4.2 Extraction Results

Owing to the lack of labels, it is hard for us to evaluate our method automatically. So we use human evaluation to measure the performance of this method. Specifically, the accuracy of our method in testing data is 0.64. The figure.3 shows the part extraction results of our method with sentences and their POS sequence input. We can find that this method can partly solve the problem of extracting terminologies from unlabelled text. Also, it has a good performance on method words. However, when it comes to very professional terminologies, the performance may be lower.


**Figure 3.** Part Results

## 5 Conclusion

In order to extract terminologies from scientific texts, we propose a unsupervised method based on sentence pattern and POS sequence of sentence. This method can extract terminologies without learning on labelled data and just need a few initial sentence patterns to cool start. Then it can learn new patterns and POS sequences on unlabelled data. After that, we can use these patterns and POS sequences to extract new terminologies from new scientific sentences. Experiments on paper abstract sentences from Web of Knowledge show that our method can achieve a 0.64 accuracy on our testing data and are practically useful on unlabelled data extraction.

## References


[1] Hua Bolin. 2013. Extracting Information Method Term from Chinese Academic Literature. *New Technology of Library and Information Service* 6 (2013), 68–75.

[2] Zhao Dongyue, Du Yongping, and Shi Chongde. 2018. Scientific Literature Terms Extraction Based on Bidirectional Long Short-Term Memory Model. *Technology Intelligence Engineering* 4, 1 (2018), 67–74.

[3] Xu Hao, Zhu Xuefang, and etc Zhang Chengzhi. 2019. System Analysis and Design for Methodological Entities Extraction in Full Text of Academic Literature. *Data Analysis and Knowledge Discovery* 3, 10 (2019), 29–36.

[4] Sun le, Jin Youbing, and etc Du Lin. 2000. Automatic Extraction of Bil ingual Term Lexicon from Parallel Corpora. *Journal of Chinese Information Processing* 14, 6 (2000), 33–39.

[5] Liu Li and Xiao Yingyuan. 2017. A statistical domain terminology extraction method based on word length and grammatical feature. *Journal of Harbin Engineering University* 38, 9 (2017), 1437–1443.

[6] Yu Li, Qian Li, and etc Fu Changlei. 2019. Extracting Fine-grained Knowledge Units from Texts with Deep Learning. *Data Analysis and Knowledge Discovery* 1 (2019), 38–45.

[7] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[8] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075* (2017).

[9] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[10] Wang Miping, Wang Hao, and etc Deng Sanhong. 2016. Extracting Chinese Metallurgy Patent Terms with Conditional Random Fields. *New Technology of Library and Information Service* 6 (2016), 28–36.

[11] Serhan Tatar and Ilyas Cicekli. 2011. Automatic rule learning exploiting morphological features for named entity recognition in Turkish. *Journal of Information Science* 37, 2 (2011), 137–151.

[12] Zeng Wen, Xu Shuo, and etc Zhang Yunliang. 2014. The Research and Analysis on Automatic Extraction of Science and Technology Literature Terms. *New Technology of Library and Information Service* 1 (2014), 51–55.

[13] Tan Ying and Tang Yifei. 2020. Automatic Extraction of Factual Knowledge Element from Scientific Literature. *Information Science* 4 (2020), 4.

[14] Lin Zefei and Ou Shiyan. 2019. Research on Chinese Named Entity Linking Based on Multi-feature Fusion. *Journal of the China Society for Scientific and Technical Information* 38, 1 (2019), 68–78.