An Unsupervised Method for Terminology Extraction from Scientific Text

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

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

Qiang Ma
Department of Information Management, Peking University

Jiaying Liu
Department of Information Management, Peking University

Hongwei He
Department of Information Management, Peking University

Keqi Chen
Department of Information Management, Peking University

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

Keywords: terminology extraction, unsupervised method, scientific text

1 Introduction

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

To overcome this problem, we propose an 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. 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 can obtain new patterns and get more terminologies by using new patterns. After several iterations, most terminology in scientific sentences can be extracted.

2 Related Work

Recent years, terminology extraction has attracted more and more attention. And all kinds of methods are produced. Some methods rely on string, syntax and other original features. Liu li[2] and Zen Wen[8] use length of word and grammatical features to choose terminology candidates. Nowadays, some methods based on machine learning and deep learning are put forward. Among these methods, LSTM[1] and CRF[6] and their variants achieve the best performance. However, they rely on labelled data and have a poor performance on new unlabelled data. To solve this problem, some semi-supervised and unsupervised methods are proposed. A graph-based semi-supervised algorithm[4] achieve a high F1 on SemEval Task 10. Automatic rule learning based on morphological features method[7] is used to extract entities without annotated data. However, owing to the difficulty of searching optimal parameters, these methods can't get fully developed.

3 Method

3.1 Overview

Our method aims to extract terminology from unlabelled data. For this purpose, we utilize two features of terminology: surrounding words and POS sequences. The process can be divided into two steps. One step is to cold-start model with unlabelled data. In this step, the model will get sentence patterns, POS sequences of terminology from data. Another step is to extract terminology with POS sequences and sentence patterns learned by model. For a sentence, the model can extract terminology with learned sentence pattern or POS sequences.

3.2 Sentence Patterns

```
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 Examples

Our sentence pattern is represented by regular expression. 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. Boundary words like "by, to, for" are used to limit the range of terminology words. What we want is matched by "(.+?)". When generating new patterns, we can use words from matched

sentence to replace the extendable part of extant pattern. For examples in figure.1, the extendable parts are "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.

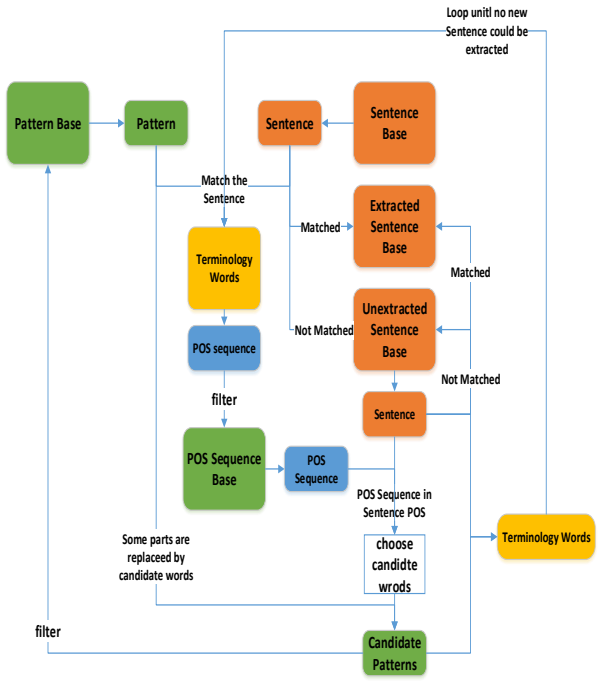

Figure 2. Cold Start Process

3.3 Cold Start

The process of cold start of our method is shown in figure.2. The inputs are sentences and their POS sequences and form the sentence base. First, we use each pattern from pattern base to match each sentence from sentence base. At beginning, pattern base only contains initial sentence patterns. Matched 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. The two bases are empty before . 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 from 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, 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 from New Data

After cold start, 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 test our method, we crawled 200k+ abstracts from Web of Knowledge. Their topics include machine learning, big data and data mining. We utilize nltk[3] to split abstracts into sentences and splitted sentences into tokens. Also we use stanfordnlp[5] to get POS tags and dependency relations of cut sentences. Our method only needs to use the tokenized sentences of abstracts and their POS tags.

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, we use human evaluation to measure our method's performance. We use training data to cold-start our model and extract 146902 terminologies from training and testing data. Specifically, the accuracy of our method in testing data is 0.64. According to some cases of result, we can find that this method can partly solve the problem of extracting terminologies from unlabelled texts. However, when it comes to very professional terminologies, the performance may be lower.

5 Conclusion

To extract terminologies from scientific texts, we propose an 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 cold-start. Then it can learn new patterns and POS sequences on unlabelled data and use them to extract new terminologies. In the future, we will test our model on standard datasets and compare it with some baselines.

References

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.