An 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

Qiang Ma
maqiang@pku.edu.cn
Department of Information Management, Peking University
Beijing, China

Jiaying Liu
ljying@pku.edu.cn
Department of Information Management, Peking University
Beijing, China

Hongwei He
1700016613@pku.edu.cn
Department of Information Management, Peking University
Beijing, China

Keqi Chen
1800016631@pku.edu.cn
Department of Information Management, Peking University
Beijing, China

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

Keywords: terminology extraction, unsupervised method, scientific text

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

1 Introduction

With the 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 (word and phrase) from papers automatically becomes an important problem.

Generally, finding new terminology relies on named entity recognition(NER). 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 are 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 obtain new patterns and are able to use them to get more terminologies. After several iterations, most terminology in scientific sentences can be extracted.

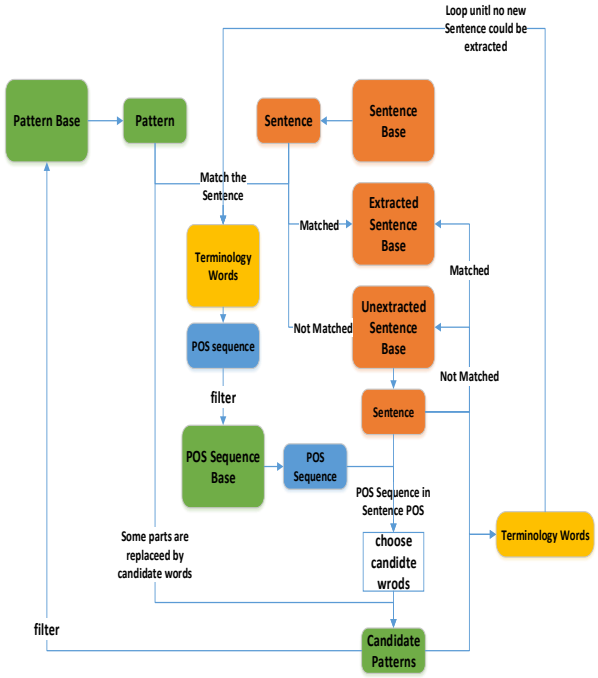

Figure 2. Cold Start Process

2 Method

2.1 Overview

Our method aims to extract terminology from unlabelled scientific texts. 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 cold start model with unlabelled data. In this step, the model will learn sentence patterns, POS sequences of terminology from data. Another step is to extract terminology with learned model. For a sentence, the model can extract terminology with learned sentence pattern or POS sequences.

2.1.1 First Step. First, we should obtain POS sequences of sentences. Then a few patterns are used to match sentences. These patterns are specially designed regular expressions. When patterns 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 not matched. Once sentence's POS sequences contains terminology's POS sequences, the sentence is matched. Then we use certain tokens in sentences to replace the extendable part in patterns to produce new patterns. After that, new found patterns will be used to match not matched sentences. After several iterations, we can get new patterns and terminologies and their POS sequences.

2.1.2 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.

2.2 Pattern Description

Pattern is a kind of 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. Border 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.

2.3 Cold Start

The process of cold starting of our method is shown in figure.2. The input of the method are sentences and their POS sequences. First, we use each pattern from initial pattern base to match each sentence from sentence base. If matched, 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, 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.

2.4 Extraction from New Data

After model 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.

3 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. After that, we can use these patterns and POS sequences to extract new terminologies from new scientific sentences. In the future, we will test our model on standard datasets and compare it with some baselines.

References