Research Article Open Access

Wei Shao, Bolin Hua*, Linqi Song

# A Pattern and POS Auto-Learning Method for Terminology Extraction from Scientific Text

https://doi.org/10.2478/dim-2021-0005
received November 11, 2020; accepted February 18, 2021.

**Abstract:** A lot of new scientific documents are being published on various platforms every day. It is more and more imperative to quickly and efficiently discover new words and meanings from these documents. However, most of the related works rely on labeled data, and it is quite difficult to deal with unlabeled new documents efficiently. For this, we have introduced an unsupervised method based on sentence patterns and part of speech (POS) sequences. Our method just needs a few initial learnable patterns to obtain the initial terminology tokens and their POS sequences. In this process, new patterns are constructed and can match more sentences to find more POS sequences of terminology. Finally, we use obtained POS sequences and sentence patterns to extract terminology terms in new scientific text. Experiments on paper abstracts from Web of Knowledge show that this method is practical and can achieve a good performance on our test data.

**Keywords:** auto-learning, terminology extraction, unsupervised method, scientific text

## 1 Introduction

With the rapid development of science and technology, more and more papers are being produced every day. So it is more and more difficult for researchers to discover something new by reading papers. Automatic term recognition (also known as term extraction) is a crucial component of many knowledge-based applications, such as automatic indexing, knowledge discovery, terminology mining and monitoring, knowledge management, and so on (Maynard, Li, & Peters, 2008). The ways and means of finding new terminologies from recently published papers instantly with the help of a computer become significant problems. Generally, finding a new terminology relies on named entity recognition (NER). However, many high-performance methods need the support of labeled data (Mintz, Bills, Snow, & Jurafsky, 2009). Although they can obtain excellent results on training and testing data, it is hard for them to process new unlabeled data that we often face. One factor responsible for this gap is that the new scientific document text features are different from the features on learning models with training data, and this is due to the difference between their domains. Also, these new scientific texts usually lack labels for extraction. So an unsupervised method that can also adapt to different fields is needed.

To overcome this difficulty, we propose a pattern and POS auto-learning method. In detail, we initialize a few patterns to extract terminologies in some sentences. In this step, we can obtain some terminologies and their POS sequences with some natural language processing tools [NLTK (Bird, 2006), StanfordNLP (Manning et al., 2014), 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 particular words in this sentence to replace the extendable parts of initial patterns. In this case, we obtain new patterns and can use these new patterns to match other sentences to get more terminologies. After several iterations, plentiful terminologies in scientific sentences can be extracted. The result shows that we can get high performance on unlabeled texts from paper abstracts from Web of Knowledge.

In summary, we propose a pattern and POS auto-learning method for terminology extraction from scientific texts, which partly solve the difficulty of extracting from unlabeled data in different fields. Experiments show that our approach can achieve a level of 0.58 precision, 0.65 recall, and 0.61 F1 score on our test data from Web of Knowledge.

***Corresponding author: Bolin Hua,** Department of Information Management, Peking University, Beijing, China.
E-mail: huabolin@pku.edu.cn
**Wei Shao, Linqi Song,** Department of Computer Science, City University of Hong Kong, Hong Kong, China

## 2 Related Work

In recent years, terminology extraction has attracted more and more attention. Several methods are being introduced to achieve better performance. Some methods rely on the string, syntax, and other original features. Bosma aims to create a semantic model of a domain to find the domain's complete terminology, consisting of terms and relations such as hyponymy and meronymy and connected to generic wordnets and ontologies (Bosma & Vossen, 2010). Shah proposes a novel similarity-driven learning approach for automatic terminology extraction for the materials science domain. They use various intra-domain and inter-domain unsupervised corpus level features to score and rank candidate terminologies (Shah, Sarath, & Reddy, 2019). Mooney presents a method to mine rules from a database extracted from a corpus of texts that are used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system (Mooney & Nahm, 2004). Liu and Xiao (2017) and Zen et al. (2014) use the length of the word and the grammatical features to choose terminology candidates. External resources such as dictionary (Tan & Tang, 2020), lexicon (Hua, 2013), parallel corpora (Sun, Jin, Du, & Sun, 2000), Wikipedia (Lin & Ou, 2019), and so on are often used to extract terminologies from unlabeled data. However, these methods have a low performance when they deal with domains with low or no resource supports.

With the development of deep learning and machine learning, some machine learning methods are put forward. Zhan presents a new terminology extraction approach combining machine learning based on cascaded conditional random fields (CRFs) with a corpus-based statistical model. In this approach, first, the low-layer and high-layer CRFs are used to extract the simple and compound terminologies, respectively (Zhan & Wang, 2015). Ha¨tty proposes two novel models to exploit general- *vs*. domain-specific comparisons: a simple neural network model with pre-computed comparative-embedding information as input and a multi-channel model computing the comparison internally. Both models outperform previous approaches, with the multi-channel model performing at the optimum level (Ha¨tty, Schlechtweg, Dorna, & Im Walde, 2020). Among these methods, Long-Short Term Memory Network (LSTM) (Zhao, Du, & Shi, 2018) and CRF (Wang, Wang, Deng, & Wu, 2016) and their variants achieve the best performance. Although these methods can obtain better results, they rely on a large number of labeled data and have a poor result on new unlabeled data.

Some semi-supervised and unsupervised methods are proposed to bridge the gap between training data and practical data. A graph-based semi-supervised algorithm (Luan, Ostendorf, & Hajishirzi, 2017) 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 the morphological features method (Tatar & Cicekli, 2011) is also used to extract entities that do not need any annotated data. However, owing to the difficulty of searching optimal parameters, these methods cannot get fully developed. Besides a single method or algorithm, terminology extracting systems utilizing all kinds of ways are practical in the real world are also focused on by researchers. Xu, Zhu, and Zhang (2019) put forward an extracting system scheme based on traditional processing. Yu, Qian, Fu, and Zhao (2019) designed a system that uses seed terminology words from a scientific database to create annotated data and train the deep learning extracting model with these labeled data. Both these systems can achieve high accuracy or recall.

In summary, compared with supervised methods, unsupervised methods can extract terminology entities from text without annotations. Also, supervised models can achieve a good result on training and testing datasets, but they may suffer a significant loss in performance when used to process new data. And unsupervised models can be more adaptive when facing new texts. Since high-performance methods find it hard to deal with unlabeled text, and many unsupervised methods rely on external resources, we propose an unsupervised method based on POS sequences and sentence patterns to extract terminology entities from scientific text. This work is also the beginning of a cold start extraction system.

## 3 Method

### 3.1 Overview

Our method aims to extract terminology from unlabeled scientific texts. For this purpose, implicit features of sentences are taken into full consideration. In detail, we utilize two features of terminology. One is the surrounding words, and another is the POS sequences of terminology.

Our method's extraction process can be divided into two steps. The first step is to cold start our model with unlabeled data. In this step, the model will learn sentence patterns, POS sequences of terminology from the input data. Besides the sentence text, each token's POS in the sentences is also needed in this step. The second step is

```
r"(.+?) (?:is|was|are|were)proposed (?:by|to|for|with|that)", 1, "proposed"
r"(?:we|to|and|then|here) (?:propose|proposed)(.+?) (?:by|to|for|with|that)"
```

***Figure 1.*** Sentence pattern cases.

to extract terminology with learned sentence patterns and POS sequences. For a new sentence, the model can extract terminology with sentence patterns when only sentence string is input. The model can also use POS sequences to extract terminology entities if the sentence's POS sequence is also input with sentence string.

Next, this paper will introduce the pattern we use, the way this method cold starts, and the way it extracts terminology from new data after a cold start.

## 3.2 Sentence Pattern

The sentence pattern used to extract target entities from sentence string 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 that often appears with method words at the same time. Border words like "by," "to," and "for" are used to limit the range of terminology words. What we want is matched by (.+?).

For the generation of new patterns, we can use words from matched sentences to replace the extendable parts of extant patterns. For patterns one and two in Figure 1, the extendable part is "propose" and "proposed." They can be replaced by "develop," "present," "put forward," and so on. In detail, we select these candidate words by their POS tags. For this pattern, words whose POS tags are "V" are chosen to replace extendable parts whose POS tags are also "V" to generate new patterns. In this case, new patterns are obtained and can be used to extract terminology in other sentences.

## 3.3 Cold Start

In this part, our method needs to learn patterns and POS sequences for terminology extraction from the input data. To obtain the input data, first, we use tools to get POS sequences of sentences. Then the initial patterns are used to match sentences. These patterns are specially designed regular expressions consisting of special words and matching groups. When the pattern matches a sentence, we can get terminology string from matched groups. After filtering and post-processing, suitable terminology tokens and their POS sequences are received as output.

Next, POS sequences of extracted terminology tokens are used to match sentences that are not matched. Once POS sequences of a sentence contain these extracted terminology POS sequences, the sentence could be regarded as a matched sentence. Then we use specific tokens in sentences to replace the extendable part in patterns to produce new patterns. The detailed processes have been described in the Pattern Description part. After that, newly found patterns are used to match not matched sentences continuously. After several iterations, we can get more patterns and terminologies and their POS sequences for extraction.

The detailed process of cold starting our method is shown in Figure 2. The inputs are sentences, including their POS sequences from scientific texts. First, we use each pattern from the initial pattern base to match each sentence from the 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 the 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 the POS sequence base. Then, for each sample of the POS sequence base, we need to find if the POS sequence of the sentence in the unextracted sentence base contains this sample. If it includes, we can choose the candidate words from this sentence to generate new patterns. After new patterns are generated, we use them to match sentences in the unextracted sentence base to obtain new terminologies. Then, we can filter new generated patterns according to their matching results and move suitable patterns to the pattern base. Next, new terminology words replace the initial extracted terminology words to participate in the extraction loop until no new sentence could be extracted.

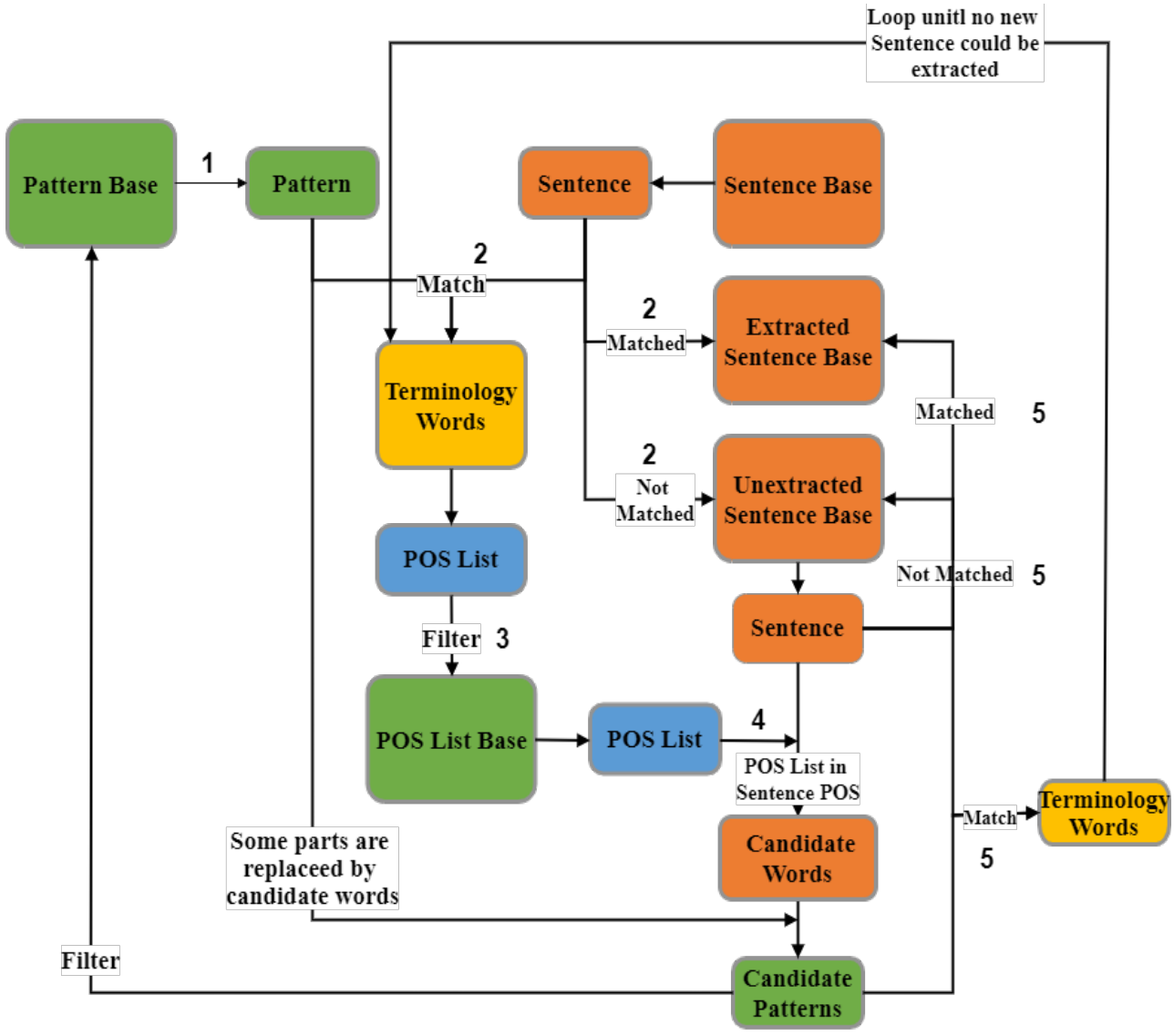


*Figure 2.* Cold start process. The numbers indicate the order of steps.

## 3.4 How to Filter

### 3.4.1 Filter of POS Sequences

After obtaining terminology words by matching sentences with patterns, we should clean these results to output suitable POS sequences for more accurate extraction. In detail, we should discard some tokens which contain little valid information. To achieve this goal, for the given terminology words, we retain the POS token, which is "DT" or "JJ" or "NN" or "CD" or "VBG" or "VBN" or "VBP" or "VBD". Other tokens will be discarded. Then, we only select POS sequences whose tokens' indexes are continuous after the filter.

### 3.4.2 Filter of Sentence Patterns

For constructed sentence patterns, we need to evaluate their quality and discard ones with low quality. We use the number of sentences matched by the newly built sentence patterns as the key metric for this purpose. In detail, we use new generated sentence patterns to extract sentences and check if the number of matched sentences exceeds the certain percentage for all sentences. If the number of matched sentences is larger than the threshold, we conserve this sentence pattern. Otherwise, we should discard it.

### 3.5 Extraction for New Data

After the method cold starts on unlabeled data, we can obtain sentence patterns and POS sequences of terminology words. Here are two approaches to get new terminologies from new unlabeled 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 the 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 the Web of Knowledge. These abstracts are from different domains, including machine learning, big data, and data mining.

As for preprocessing, we utilize the NLTK to split abstracts into sentences and split sentences into tokens. Also, we use the 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, and dependency relations of sentences. In this method, we use the tokenized sentences of abstracts and POS tags of sentences.

In the experiment, we use 54k+ sentences and their POS sequences as training data without labels and 500 sentences and their POS sequences as test data with labels. A labeled sentence consists of a token sequence and a label sequence. When the token is a terminology word, the corresponding label is 1. Otherwise, the label is 0. To ensure the accuracy of annotation, we manually annotate 500 sentences in the test set.

### 4.2 Results on Our Dataset

We use recall, precision, and F1 score as the metrics to compare our method with two rule-based methods on our dataset. For unsupervised learning methods, its extraction results are usually full of noisy terms. If it is evaluated by a hard standard, the performance may be very low so that it is difficult to compare these unsupervised methods because their performances are close. So MUC-6 (Grishman & Sundheim, 1996) propose a relaxed-match evaluation, a soft standard used to check if an extraction result is correct. In detail, if an extraction result has an overlap with the real terminology, this extracted term is regarded as a valid result.

**Table 1**
*The Performance of Our Method*

| Method | Precision | Recall | F1 score |
|---|---|---|---|
| Our method | 0.58 | 0.65 | 0.61 |
| Rule-based method 1 | 0.55 | 0.54 | 0.54 |
| Rule-based method 2 | 0.52 | 0.42 | 0.46 |

Related data and codes could be found in https://github.com/visionshao/TerminologyExtraction.

In our experiments, if the number of words appearing in an extracted term and a terminology term in this sentence at the same time exceeds the product between the number of words in the terminology term and a cover percentage (this number is set as 0.65 in our experiments), we come to a conclusion that this extracted term is correct. Following this setting, we obtain our method and other rule-based methods' performance, which is shown in Table 1.

For rule-based method 1, it uses continuous nouns as the terminology term. Rule-based method 2 chooses to filter non-noun and non-verb words and use left words with continuous indexes in sequence before filtering as the terminology term. According to Table 1, our method outperforms the other two methods on all metrics.

### 4.3 Case Study

Figure 3 shows four extraction cases (s1, s2, s3, and s4) of our method with sentences and their POS sequence input. Each case contains a sentence and an extraction result. The sentence consists of several words and forms a list of words. And terminology words are shown in blue in each sentence. The list of the word list is the extraction result. The correct terminology term (a word list) is also in blue. According to these four cases, we can find that this method can partly solve the problem of extracting terminologies from the unlabeled text, such as case 1 and case 2. Also, it has a good performance on method words. However, as shown in case 4, the performance may be lower when it comes to very professional terminologies (ShiftASA, chemical-shift in case 4).

S1

['Giving', 'the', 'highest', 'classification', 'accuracy', ',', 'support', 'vector', 'machine', 'technique', 'outperformed', 'the', 'others', 'with', 'a', 'value', 'of', '78.83', '%', '.']

[['the', 'highest', 'classification', 'accuracy'], ['support', 'vector', 'machine']]

S2

['The', 'preliminary', 'experimental', 'results', 'demonstrate', 'that', 'our', 'developed', 'system', 'is', 'workable', ',', 'allowing', 'for', 'prediction', 'of', 'possible', 'evolution', 'and', 'early', 'warning', 'of', 'critical', 'incidents', 'with', 'a', 'support', 'of', 'dynamic', 'entity', 'extraction', '.']

[['preliminary', 'experimental', 'results'], ['dynamic', 'entity', 'extraction']]

S3

['Present', 'proof-of-concept', 'study', 'shows', 'that', 'modelling', 'of', 'multiple-source', 'geochemical', 'soil', 'data', 'using', 'machine-learning', 'algorithms', 'can', 'be', 'successfully', 'accomplished', 'and', 'that', 'model', 'predictions', 'nicely', 'complement', 'current', 'interpretation', 'and/or', 'established', 'archeological', 'predictive', 'modelling', 'of', 'areas', 'of', 'archaeological', 'interest', '.']

[['Present', 'proof-of-concept', 'study'], ['multiple-source', 'geochemical', 'soil'], ['archeological', 'predictive', 'modelling'], ['using', 'machine-learning', 'algorithms']]

S4

['Using', 'machine', 'learning', 'techniques', 'we', 'developed', 'an', 'algorithm', 'called', '"', 'ShiftASA', '"', '"', 'that', 'combines', 'chemical-shift', 'and', 'sequence', 'derived', 'features', 'to', 'accurately', 'estimate', 'per-residue', 'fractional', 'ASA', 'values', 'of', 'water-soluble', 'proteins', '.']

[['an', 'algorithm', 'called'], ['machine', 'learning', 'techniques'], ['per-residue', 'fractional', 'ASA'], ['Using', 'machine', 'learning']]

***Figure 3.*** A few cases.

## 5 Conclusion

To extract terminologies from scientific texts, we propose a cold start method based on sentence pattern and POS sequence of the sentence. This method can extract terminologies without learning on labeled data and just need a few initial sentence patterns to a cold start. Then it can learn new patterns and POS sequences on unlabeled data. We can then 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 0.58 precision, 0.65 recall, and 0.61 F1 scores on our test data when the cover percentage is 65%, which shows that our approach is practically useful for unlabeled data extraction.

## References

Bird, S. (2006). NLTK: The natural language toolkit. *ArXiv, cs.CL/0205028*. doi: 10.3115/1225403.1225421

Bosma, W., & Vossen, P. (2010). Bootstrapping language neutral term extraction. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23*.

Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 466-471. doi: 10.3115/992628.992709

Ha¨tty, A., Schlechtweg, D., Dorna, M., & Im Walde, S. S. (2020). Predicting degrees of technicality in automatic terminology extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2883-2889. doi: 10.18653/v1/2020.acl-main.258

Hua, B. (2013). Extracting information method term from Chinese Academic Literature. *New Technology of Library and Information Service*, 6, 68-75. doi: 10.11925/infotech.1003-3513.2013.06.11

Lin, Z., & Ou, S. (2019). Research on Chinese named entity linking based on multifeature fusion. *Journal of the China Society for Scientific and Technical Information*, *38*(1), 68–78.

Liu, L., & Xiao, Y. (2017). A statistical domain terminology extraction method based on word length and grammatical feature. *Journal of Harbin Engineering University*, *38*(9), 1437-1443. doi: 10.11990/jheu.201605037

Luan, Y., Ostendorf, M., & Hajishirzi, H. (2017). Scientific information extraction with semi-supervised neural tagging. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2641–2651. doi: 10.18653/v1/D17-1279

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. doi: 10.3115/v1/P14-5010

Maynard, D., Li, Y., & Peters, W. (2008). NLP techniques for term extraction and ontology population. *Ontology Learning and Population*. Retrieved from https://gate.ac.uk/sale/olp-book/main.pdf

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Volume 2 - Volume 2)*, 1003-1011. doi: /10.5555/1690219.1690287

Mooney, R. J., & Nahm, U. Y. (2004). Text mining with information extraction. *Proceedings of the 4th International MIDP Colloquium*, 141-160.

Shah, S., Sarath, S., & Reddy, S. (2019). Similarity driven unsupervised learning for materials science terminology extraction. *Computacio´ny Sistemas*, *23*(3), 1005-1013. doi: 10.13053/CyS-23-3-3266

Sun, L., Jin, Y., Du, L., & Sun, Y. (2000). Automatic extraction of bilingual term lexicon from parallel corpora. *Journal of Chinese Information Processing*, *14*(06), 33-39.

Tan, Y., & Tang, Y. (2020). Automatic extraction of factual knowledge element from scientific literature. *Information Science*, 4, 23-27.

Tatar, S., & Cicekli, I. (2011). Automatic rule learning exploiting morphological features for named entity recognition in Turkish. *Journal of Information Science*, *37*(2), 137–151. doi: 10.1177/0165551511398573

Wang, M., Wang, H., Deng, S., & Wu, Z. (2016). Extracting Chinese metallurgy patent terms with conditional random fields. *Data Analysis and Knowledge Discovery*, *32*(6), 28–36. doi: 10.11925/infotech.1003-3513.2016.06.04

Xu, H., Zhu, X., & Zhang, C. (2019). System analysis and design for methodological entities extraction in full text of academic literature. *Data Analysis and Knowledge Discovery*, *3*(10), 29–36.

Yu, L., Qian, L., Fu, C., & Zhao, H. (2019). Extracting fine-grained knowledge units from texts with deep learning. *Data Analysis and Knowledge Discovery*, *3*(1), 38–45. doi: 10.11925/infotech.2096-3467.2018.1352

Zeng, W., Xu, S., Zhang, Y., & Zhai, J. (2014). The research and analysis on automatic extraction of science and technology literature terms. *New Technology of Library and Information Service*, 1, 51–55.

Zhan, Q., & Wang, C. (2015, July). A hybrid strategy for Chinese domain-specific terminology extraction. Paper presented at the *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland. doi: 10.1109/IJCNN.2015.7280490

Zhao, D., Du, Y., & Shi, C. (2018). Scientific literature terms extraction based on bidirectional long short-term memory model. *Technology Intelligence Engineering*, *4*(1), 67–74.