Research on Text Mining of Material Science Based on Natural Language Processing

Xiang Gao\textsuperscript{1,}\textsuperscript{*}, Rong Tan\textsuperscript{2} and Guanghui Li\textsuperscript{1}

\textsuperscript{1}School of Computer Science, Northwestern Polytechnical University, Xian, China
\textsuperscript{2}School of Software Engineering, Northwestern Polytechnical University, Xian, China

*Corresponding author e-mail: gaoxg@nwpu.edu.cn

Abstract. Facing large-scale and rapidly growing material science literature data, text mining has become a research hotspot of material science. In recent years, natural language processing technology and machine learning methods have become the main technical means of text mining in materials science. The main task of text mining is to transform unstructured text data into structured material data by information extraction methods such as Named Entity Recognition and entity relationship extraction. This research proposes a general solution framework for material information extraction tasks, and introduces the main concepts and processes of text processing, text annotation, entity relationship extraction, etc., and discusses the current research progress and possible future research directions.

1. Introduction

A large number of papers, patents and journals form a huge unstructured information corpus in the field of materials science. In the face of large-scale and rapidly growing material science literature data, even experts in the field can not rely on manual methods to obtain useful information, so as to fully grasp the research status and future development trend in this field. [1]. Text is unstructured and semi-structured information represented by natural language which is hard to interpret by computer. [1]. Manual information extraction is poorly accurate and requires a lot of time and resources [1]. Therefore, the need for text mining technology to quickly and efficiently extract material science knowledge from the literature database has become very urgent.

Text mining refers to the method of extracting comprehensible and useful information from text corpus. Text mining is a direction of data mining, and its objects are unstructured or semi-structured documents. In recent years, text mining of materials science mainly relies on natural language processing technology and machine learning methods, which can effectively extract the required information and discover the hidden material science knowledge from a large number of materials science documents [2]. As shown in Figure 1, text mining compiles unstructured text data into structured material data and stores it in structured database. The research of text mining in materials science will improve the efficiency of building material gene database, accelerate the pace of data-driven and first-principles materials design [1].
2. Task description and solution framework

2.1. Task description

The research goal of this research is to extract material information from literature by using text mining method and automatically generate structured material database. This research mainly depends on natural language processing technology and machine learning method, combining custom rules and professional noun dictionary. Material entity recognition (MER) and entity relationship extraction are the core tasks and important links of material information extraction.

Material entity recognition refers to finding and classifying predefined entities in text, usually in different forms of compounds [2]. Entity relation extraction needs to extract predefined entity relation from unstructured text on the basis of entity recognition. The relationship between entities can be formalized as a relation triplet \( \langle e_1, r, e_2 \rangle \). In the triplet, \( e_1 \) and \( e_2 \) represent entities, and \( r \) belongs to the target relation set \( R = \{r_1, r_2, r_3, ..., r_i\} \). The task of entity relation extraction is to extract relation triplet \( \langle e_1, r, e_2 \rangle \) from natural language text, so as to extract useful material information [2].

2.2. Solution framework

For the text mining task of material documents, the framework of material information extraction based on natural language processing is shown in Figure 2. The following are common operations in text mining tasks.

1. Document processor. It need to analyze PDF, HTML, XML document structure, and separate picture, text and table. The recognition of image information needs the support of optical character recognition technology (OCR), which is not in the scope of this study. It is often necessary to convert file formats and extract plain text from different files.

2. Text preprocessing: text cleaning, sentence splitting, tokenization, spell checker, pos tagging, lemmatization, stemming, case folding, etc.

3. Rules and dictionaries: self defined matching rules and professional noun dictionary (entity name, relation noun, etc.). Regular expression matching is a simpler method for entities that can be easily identified by rules.

4. Constructing word vector. The labeled sentence is segmented, and every word is encoded as a word vector acceptable to computers. Find out the relative position of each word and the entity pair in the sentence as the position vector of this word. The combination of word vector and position vector is regarded as the final vector representation of the word.

5. Entity boundary detection: using CRF, HMM machine learning method or CNN, LSTM and other deep neural networks to identify named entities.

6. Relationship classification. According to the pre-defined relationship types, the feature vector of sentences is put into the non-linear layer to classify, and the final relationship between entities is extracted.

7. Evaluate classification performance: evaluate the relationship classification results.
3. Text mining in Materials Science

3.1. Text preprocessing

3.1.1. Document segmentation. For the text mining task of material documents, it is impractical to use the full text regardless of the format and content of the original document. Thus document segmentation is needed to deal with documents more effectively. Documents in markup languages such as HTML are usually easier to handle because they provide formatting tags. For PDF document segmentation, regular expressions are generally used. The flow chart of document segmentation is shown in Figure 3. Abstract, full text, caption and tables usually contain very different information. For example, the abstract part of the document often contains the name of the main material entity, and the experimental part often contains the description of the material experiment process. It is more efficient to divide the document into different parts and extract different material information [3]. The tables in the document are also an important source of experimental data. Table mining is a main research direction of text mining. How to accurately locate and identify structured forms in complex document structures is an important issue [4].

3.1.2. Sentence boundary disambiguation. Natural language is hierarchical, with words forming phrases, phrases forming sentences, and sentences forming chapters. From a fine-grained perspective, sentences as the direct upper-level structure of words and phrases are mostly basic objects of natural language processing, such as grammatical analysis and part-of-speech tagging [5]. After the full text is divided into sentence sequences, each sentence is taken as the processing unit. The text mining methods of materials science that have been proposed now mostly use sentences as the basic
processing unit. Sentence boundary disambiguation (SBD) is the basis of material science text mining by identifying sentence boundaries and segmenting input text into sentence sequences. Its accuracy directly affects performance of natural language processing and information extraction [2].

3.1.3. Tokenizers and Stop words. Tokenization, also known as word segmentation, is the basis of natural language processing. Its accuracy determines the quality of word vectors and text analysis. It usually needs to segment each sentence, and then extract the main components of the sentence, such as punctuations, component words, expression sequences, numbers and so on. In text preprocessing, it is the most critical step [2]. At present, the most common English word segmentation tool is Natural Language Toolkit (NLTK). There are many word segmentation tools commonly used in Chinese, such as Jieba. In order to improve the efficiency of text processing, some words without actual meaning need to be filtered out in the process of word segmentation. These words are called stop words, such as pronouns, qualifiers, prepositions, function words and so on. Removing them does not affect the semantic integrity of the sentence. In English, articles, prepositions and pronouns such as "a", "the", "to", "their" are all stop words. We can directly use the English stop words list provided in NLTK to remove these stop words.

3.2. Text annotation
Most tasks of natural language processing are supervised learning. For the problems of sequence annotation and classification, such as Named Entity Recognition and relationship recognition, the model training of annotation data is needed. For some vertical fields, such as materials, medicine and finance, there are many proper terms and special requirements. So the model trained on the general corpus (e.g. Wikipedia) cannot be used directly. Labeling and managing training datasets by hand is one of the biggest bottlenecks in this task [1]. The process of traditional manual data annotation is often tedious and inefficient. At present, some open-source integrated visual text annotation tools have appeared in the field of machine learning. They have friendly annotation interface, which makes the annotation operation as simple and intuitive as possible, and greatly improves the efficiency of the project [4]. The user writes the custom rules and dictionaries to the configuration file, adds data annotation in the interactive GUI, and automatically generates formatted annotation data, such as JSON, SQL, etc. Table 1 is a comparison of several popular open source text annotation tools.

| Text annotation tools | Platform | Description |
|-----------------------|----------|-------------|
| IEPY                  | Web      | A web-based information extraction tool mainly used for relationship extraction. |
| Mind tagger           | Web      | Supporting output in SQL, CSV, TSV and JSON formats. |
| SUTD Annotator        | Windows  | A lightweight corpus annotation tool supporting multiple languages. |
| Prodigy               | Multi-platform | A new tool for radically efficient machine teaching. |
| BRAT                  | Web      | Open source text annotation tool based on Web. |
| Snorkel               | Multi-platform | Programmatically building and managing training data. |

3.3. Material entity recognition
In the general field, Named Entity Recognition (NER) refers to the recognition of named reference items from text, such as person name, organization name, etc., which paves the way for tasks such as relationship extraction. Within a particular domain, the various entity types within the domain are defined accordingly. In the field of materials science, named entities usually refer to the chemical entities that make up materials. Due to the diversity of writing styles, the names of chemical entities in texts are usually expressed in many different ways, such as formulas (e.g., “Al2O3”), chemical families (e.g., “halogens”), company codes (e.g., “ICI206636”), acronyms (e.g., “PEI”) and so on. Useful information such as the composition elements and partial properties of the compound can be extracted from the representation of chemical entities, but it usually does not contain the structural information of the compound [2].
Named Entity Recognition (NER) is the key and foundation of material science text mining task. At present, shallow machine learning and deep neural network are the two main methods of NER in material science. Shallow machine learning methods are mainly divided into Support Vector Machine (SVM), Maximum Entropy model (ME), Hidden Markov model (HMM), Conditional Random Field model (CRF), etc. For example, Manabu et al. fused CRF, HMM, ABNER, and other models to achieve 88.87% F-score on BiocreativeIIGM corpus [6]. However, the traditional shallow machine learning method depends on the design of artificial features to a great extent. While artificial features and domain knowledge improve the accuracy of the model, it also leads to a decline in the generalization ability of the model. Therefore, many Named Entity Recognition models based on neural network are proposed [7]. In the research of Named Entity Recognition using deep neural network, Li et al. achieved 88.6% F-score on the BiocreativeIIGM corpus by using Bi-directional LSTM [7]. Compared with rule-based method or statistical machine learning method, deep neural network method has the advantages of more generalization and less dependence on artificial features.

3.4. Entity relation extraction

Entity relation extraction is the core step of natural language processing and text mining. By modeling text information, it automatically extracts the semantic relationship between entity pairs and extracts effective semantic knowledge [7]. It specifically refers to the relationship classification of entity pairs when the named entity pairs in sentences have been recognized. The classical methods of entity relationship extraction mainly include supervision, semi-supervision, weak supervision and unsupervised. The classical methods have the problem of error propagation in feature extraction, which greatly affects the effect of entity relationship extraction. As deep learning has made great progress in information extraction research, scholars have gradually applied deep learning to entity relationship extraction tasks.

According to the different order of the two subtasks of entity recognition and relationship classification, the supervised entity relationship extraction based on deep learning can be divided into pipeline method and joint learning method. The pipeline learning method refers to the entity relationship extraction based on the entity identification has been completed. The joint learning method is based on the end-to-end model of the neural network, which can recognize entities and extract entity relationships at the same time. Zeng et al. first proposed to use CNN for relationship classification in 2014, and Katiyar et al. first combined attention mechanism with Bi-LSTM to extract entity and classify relationship in 2017 [8, 9]. The neural network model has achieved good results in the field of supervised learning.

3.5. Evaluation index

In the field of text mining, there are three basic evaluation indexes (Precision, Recall and F-score). The calculation methods of Precision and Recall are shown in equations (1) and (2), where TP represents the amount of entity pairs correctly extracted belonging to relation R, FP represents the amount of entity pairs which are mistakenly extracted as relation R, FN represents the amount of entity pairs of relation R which are wrongly extracted [10].

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (1)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (2)
\]

For relationship extraction, Recall and Precision interact with each other, and they are complementary. F-score combines the information of Precision and Recall, and its calculation formula is shown in formula (3).

\[
F_\beta = \frac{(\beta^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (3)
\]
is a parameter to adjust the proportion of Precision and Recall. In practical experiments, it is generally considered that Precision is as important as Recall. Therefore, \( \beta \) value is generally set to 1. The above formula (3) can be expressed as the following formula (4).

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

4. Conclusion

Aiming at the huge unstructured information corpus of materials science, text mining method dependent on machine learning and natural language processing algorithms has made a breakthrough. The automatic method of material information extraction can directly display the structured knowledge and accelerate the construction of material gene database. It also have great help for the design and characterization of new materials, prediction of material properties and other research. This research introduces the general solution framework and main process of natural language processing in the text mining task of materials science, summarizes text preprocessing methods such as document segmentation, sentence splitting, tokenizers, etc., compares various popular open-source integrated visual text annotation tools, and introduces the research progress of Named Entity Recognition (NER) and entity relation extraction based on natural language processing in recent years.

Material entity recognition and entity relationship extraction are the important links of material science information extraction. At present, there are still many difficulties to be overcome and research directions to be improved, such as no unified naming and expression of material entities, co-reference resolution and entity disambiguation, extraction of complex chemical structures, cross-sentence-level relationship extraction. Thus many text mining methods still rely on artificial features and domain knowledge. The purpose of text mining in materials science is to provide a good knowledge mining tool for researchers of materials science. It is also the direction of future efforts to cooperate with material science researchers with relevant background and integrate domain knowledge.

5. Acknowledgments

This work was financially supported by The National Key Research and Development Program of China (2016YFB0700504) and the Science and Technology Project of Shaanxi Province (2018GY-048).

References
[1] Ramprasad R, Batra R, Pilania G, Machine learning in materials informatics: recent applications and prospects, npj Computational Materials, 3 (2017) 54-56.
[2] Krallinger M, Rabal O, Anália, Information Retrieval and Text Mining Technologies for Chemistry, Chemical Reviews, 2017:acs.chemrev.6b00851.
[3] Court C J, Cole J M, Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction, Scientific Data, 5 (2018):180111.
[4] Joo-Chang Kim, Kyungyong Chung, Associative Feature Information Extraction Using Text Mining from Health Big Data, Wireless Personal Communications, 105 (2018) 2 1-17.
[5] Chiu J P C, Nichols E, Named Entity Recognition with Bidirectional LSTM-CNNs, Transactions of the Association for Computational Linguistics, 4 (2015) 357-370.
[6] Pinheiro P H O, Collobert R, Recurrent Convolutional Neural Networks for Scene Parsing, Journal of Machine Learning Research, 1 (2014) 82-90.
[7] Ward L, Wolverton C, Atomistic calculations and materials informatics: A review, Current Opinion in Solid State and Materials Science, 21 (2016) 3.
[8] Zeng D, Liu K, Relation classification via convolutional deep neural network, In: Proc. of the 25th Int’lConf, on Computational Linguistics: Technical Papers, 2014, pp. 2335–2344.
[9] ZHANG Runyan, MENG Fanrong, ZHOU Yong, Semantic relation extraction model via attention based neural Turing machine, Journal of Computer Applications, 2018.
[10] Zhang Q, Chen M, Liu L, A Review on Entity Relation Extraction, Control and Computer Engineering, 2017.