Application of Natural Language Processing in Environmental Protection Industry Based on Monte Carlo Tree

Zhiheng Wu
School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin, China
1801130967@qq.com

Abstract. In this paper, a method of environmental science recognition based on natural language processing model is proposed. This method takes natural language processing model as the core, adds environmental protection professional word segmentation library and external feature recognition technology on the basis of general word segmentation library, realizes automatic extraction of accurate environmental science named entity information from EIA documents, and uses MCTS Monte Carlo tree to construct search engine. The search efficiency of named entity recognition process is improved.

1. Introduction
At present, in the environmental protection industry, the number of environmental impact assessment documents is becoming larger and larger, and the degree of data integration is getting higher and higher. With the development of Internet and information technology, information management technology has been widely used in all levels and systems [1]. However, the data of environmental assessment projects contain many indicators (such as total investment, constructive nature, industry, etc). All kinds of data and information of environmental assessment projects can not be computerized, and can not be used in a timely and effective manner. The environmental statistics work is heavy and time-consuming. It is difficult to guarantee the accuracy simply by manual work. In order to reduce the labor intensity of environmental assessment staff and improve work efficiency, and implement the filing and management system of electronic documents, an entity recognition method of environmental science based on natural language processing is designed to realize intelligent extraction of environmental assessment information.

2. Theoretical basis
2.1. Environmental impact assessment
Environmental impact assessment (EIA) refers to the analysis, prediction and assessment of the possible environmental impacts after the implementation of planning and construction projects, the proposals of countermeasures and measures to prevent or mitigate adverse environmental impacts, and the methods and systems for tracking and monitoring [2]. Environmental impact assessment can provide scientific basis for decision-making of development and construction activities, for rational
distribution of economic construction, for determining the direction and scale of economic
development in a region, for formulating regional economic development plans and corresponding
environmental protection plans, for formulating environmental protection countermeasures and
scientific environmental management, and for promoting related environmental science..

2.2. Natural Language Processing
It is the technology of processing and processing various types of human-specific written and spoken
natural language information by computer tools [3]. Natural language processing, also known as
natural language processing, is one of the important contents of artificial intelligence research.

Entity (NE) refers to a specific noun (such as the name of an organization) and a specific quantifier
(time and number) of interest to people. The task of entity recognition is defined as identifying and
categorizing the distinguished names and meaningful quantitative phrases that appear in the text.
Chinese NE recognition research started late, compared with English entity recognition, Chinese NE
recognition is more difficult. Chinese text does not have identifiers like Spaces in English text to show
the boundary of the identifier, the word segmentation and entity influence each other, and it lacks
obvious characteristics. The organization is complex and irregular [4].

3. Environmental assessment information extraction method based on NLP
As shown in figure 1, the natural language processing engine is taken as the core, the document
processor is used to preprocess the environmental assessment file, and the Monte Carlo tree search
algorithm is used to extract the target information from the environmental assessment file by
combining the external feature recognizer and the professional word segmentation database of
environmental assessment. Keyword extraction is to extract some words most relevant to the meaning
of this article from the text, which has important applications.

Keyword extraction algorithm generally fall into two categories [5], supervised and unsupervised.
Supervised keyword extraction method is mainly done through the way of classification, by building a
relatively rich and perfect word table. The advantages are high precision, but the disadvantages are
that large quantities of annotation data are needed, and the glossary needs timely maintenance.

In comparison, the unsupervised method has low requirements for data. It does not need a manually
generated and maintained word list, nor does it need manually annotated corpus for auxiliary training.
At present, the commonly used keyword extraction algorithms are all based on unsupervised
algorithms. Monte Carlo tree algorithm is one of them [6].

![Algorithm of the Monte Carlo tree.](image)

Figure 1. Algorithm of the Monte Carlo tree.

When the problem to be solved is the probability of an event, or the expected value of a random
variable, they can get the frequency of the event, or the average of the random variable, by some
"trial" method, and use them as the solution to the problem. This is the basic idea of the Monte Carlo
method. Monte Carlo method, by grasping the geometric quantity and geometric characteristics of the movement of things, uses mathematical methods to simulate, that is, to carry out a digital simulation experiment. It is based on a probabilistic model, according to the process described by this model, through the simulation experiment results, as the approximate solution to the problem. The MCTS Monte Carlo tree is established through the samples of environmental assessment reports of various representative industries, and the intelligent processing model is established through the mining of these data. When the number of visits to a node exceeds a certain threshold value, a further search at the next level is carried out in the Monte Carlo tree [7].

Optimal weight selection algorithm:

\[
f(R) = \sum_{i=1}^{N} d(S_i, R(t_i))
\]

(1)

Where, \(R\) is the regular expression match, \(t\) is the text \(R(t)\) representation to be matched, the value after the regular expression match, \(s\) is the value to be matched, and function \(d\) is the editing distance. Thus, by measuring performance, enumerate all possibilities and select the maximum/minimum value to obtain the best weight.

| Table 1. MCTS Monte Carlo tree search time comparison. |
|---------------------------------|
| **Number** | **name** | **place** | **name** | **place** |
| 1 | 3.43 | 3.42 | 0.28 | 0.29 |
| 2 | 1.79 | 1.79 | 0.28 | 0.27 |
| 3 | 1.88 | 1.89 | 0.28 | 1.08 |
| 4 | 1.85 | 1.92 | 0.31 | 0.31 |
| 5 | 2.14 | 2.16 | 0.28 | 0.27 |
| 6 | 1.98 | 1.97 | 0.28 | 2.02 |
| 7 | 1.91 | 1.90 | 0.24 | 0.23 |
| 8 | 2.57 | 2.57 | 0.32 | 0.32 |
| 9 | 2.02 | 2.05 | 0.25 | 0.25 |
| 10 | 2.48 | 2.50 | 0.24 | 0.24 |
| **Average** | 2.21 | 2.22 | 0.28 | 0.53 |
| **Accuracy(%)** | 91 | 90 | 91 | 90 |

| Table 2. Contrast of Joining Professional Word Segmentation Library. |
|---------------------------------|
| **number** | **right/wrong** | **time** | **right/wrong** | **time** |
| 1 | right | 8.393 | right | 8.143 |
| 2 | right | 2.886 | right | 2.823 |
| 3 | right | 6.864 | right | 6.677 |
| 4 | right | 7.348 | right | 7.27 |
| 5 | right | 5.803 | right | 5.772 |
| 6 | right | 12.293 | right | 12.137 |
| 7 | right | 1.154 | right | 1.045 |
| 8 | right | 6.365 | right | 6.162 |
| 9 | wrong | 1.217 | right | 1.217 |
| 10 | wrong | 1.264 | wrong | 1.248 |
Then, we create external feature recognizer and introduce external features, including: dictionary of organizational structure, dictionary of unit name of pollutant, dictionary of value range of pollutant (concentration, emission, etc.), dictionary of professional terms for environmental assessment (application in word segmentation) and dictionary of national administration (provinces, cities, counties and villages).

After the MCTS Monte Carlo tree algorithm is adopted, the search time is obviously accelerated through the training of the munka chapter, as shown in table 1. This shows that MCTS provides a better method than traditional tree search. MCTS does not require any knowledge of a given domain strategy or specific practices to make a sound decision. This algorithm can work effectively without any knowledge of the game beyond the basic rules.

Also, after joining the professional word segmentation database of environmental assessment, the search accuracy is improved and the search time is shortened, as shown in table 2. This shows that after the iterative problem of MCTS is processed in advance, the search accuracy and efficiency can be improved. MCTS search may require enough iterations to converge to a good solution, which is also a problem in more general applications that are difficult to optimize.

4. Conclusion
This paper proposes an environmental science entity recognition method based on natural language processing model. This method takes natural language processing model as the core, adds environmental protection professional word segmentation database and external character recognition technology on the basis of general word segmentation database, and realizes the automatic extraction of accurate environmental science entity information from environmental assessment documents. At the same time, MCTS Monte Carlo tree is used to build the search engine, which improves the search efficiency of the entity identification process.

References
[1] Stenetorp P, Pyysalo S, Ohta T, et al. BRAT: a web-based tool for NLP-assisted text annotation [C] // Demonstrations at the Conference of the European Chapter of the Association for Computational Linguistics. 2012.
W. Strunk Jr., E.B. White, The Elements of Style, third ed., Macmillan, New York, 1979.

[2] Bontcheva K, Cunningham H, Tablan V, et al. Using GATE as an environment for teaching NLP [C] // acl-02 Workshop on Effective Tools & Methodologies for Teaching Natural Language Processing & Computational Linguistics. 2002.

[3] Bordino I, Ferretti A, Firrincieli M, et al. Advancing NLP via a distributed-messaging approach [C] // IEEE International Conference on Big Data. 2017.

[4] Markus Enzenberger M M. A lock-free multithreaded Monte-Carlo tree search algorithm [J]. 2010.

[5] Chu C Y, Hashizume H, Guo Z, et al. Combining pathfinding algorithm with Knowledge-based Monte-Carlo tree search in general video game playing [C] // Computational Intelligence & Games. 2015.

[6] Sang H H. A top-down iteration algorithm for Monte Carlo method for probability estimation of a fault tree with circular logic [J]. Nuclear Engineering & Technology, 2018, 50 (6): 854-859.

[7] Larget B, Simon D L. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees [J]. Molecular Biology & Evolution, 1999, 16 (6): 750-759.