Research of Ambiguity checking by NLP in EIA Report

Zefeng Wu¹, Peiyi Wang²*, Yanzhu Hu¹ and Jin Yu¹

¹Automation College, Beijing University of Posts and Telecommunications, Beijing, 100876, China
²Beijing Key Laboratory of Control Technology for Toxic, Hazardous, Flammable and Explosive Sources of City, Beijing Municipal Institute of Labour Protection, Beijing, 100054, China

*Corresponding author e-mail: wpy0424@sina.com

Abstract. The environment impact assessment(EIA) report is an important reference measure for measuring the overall environmental status of the region. Both the data and the text description in the report are intuitive manifestations of the environmental status. Therefore, the consistency of the text description and data is very important in the EIA report. This paper focus on text description and data ambiguity detection of atmospheric indices by using natural language processing(NLP) techniques in EIA report. The key issue of detection is matching. In the process, firstly taking the EIA of a project in Haidian District of Beijing as an example source data and establishing a corpus of atmospheric EIA. The corpus contains atmospheric environmental indicators, the degree of description of these indices, values, etc. Secondly, comparing the subjective and objective weighting method and choose a comprehensive weight method to match target for ambiguity detection. Finally, using regular expression to match the text description and data of atmospheric indices in EIA report. This method of matching can reduce the complexity compared with traditional matching method.

1. Introduction

Recently, the prevention and control of atmospheric pollution is the critical part of the governance and protection of global environment. In most countries, EIA (Environmental Impact Assessment) report is a true reflection of the atmosphere environment. A high quality EIA report is the key of governance and protection of atmospheric environment. A EIA report contains some contents like figure, table, text, data and so on. There may be ambiguity between two or more contents in EIA report. Statistical analysis shows that many ambiguity stems from the lack of rigor in text narration. Therefore, NLP (Natural Language Processing) is introduced to text analysis in EIA report.

Many researchers apply and expand the NLP method to text analysis in different fields. Sung S.F et al. abstract from specific word forms or grammatical aspects by NLP method. Then the goal of model matching is achieved [1]. They also build an automatic system to check grammatical error according to NLP method [4]. Nesi et al. apply NLP to organization addresses recognition and matching in the web domains [2]. Fredriksson et al. apply NLP to match approximate pattern in different natural language texts [3]. Badal V.D et al. identify the near-native models from a large number of the protein complexes by NLP method [5]. Afzal N et al. apply NLP method to identify the clinical records in medicine [6]. Hsu et al. establish a suitable template to eliminate the shortcomings of rule-based systems by using
NLP [7]. Sevenster et al. apply NLP to the paired measurement of the CT report [8]. Indukuri et al. present a method and a tool to do a claim similarity analysis between two different patents based on NLP method [9]. Souili et al. apply NLP method to match and extract knowledge relevant to IDM (Inventive Design Method) from patents [10]. Above studies have been done by some style about NLP technique, but less research involves in the matching of text and data by NPL. In the paper, the NPL method aims to match text and data and to check ambiguity from EIA report.

2. The NLP in the EIA Report
The text and data from the EIA report are described in natural language. Their matching can be regarded as the process of NLP. Firstly, a corpus which is suitable to EIA report should be established. Then, the subjective and objective weighting method describes the importance and frequency of vocabulary in EIA method. Lastly, the text and data is matched by the regular expression.

2.1. The Establishment of Corpus in NLP
Nowadays, the common corpuses are Chinese corpus data, Multilingual corpus, Chinese balanced corpus and so on. The general corpus is far from enough and a special corpus needs to be built. The special corpus is based on the words of EIA report. There are two principles for establishment of corpus. The first one is the truth of the statistical text rather than manual editing. The second one is using the required content. In the paper, it is necessary to establish the atmosphere related corpus and select the available figures related to the atmosphere such as $O_3$, particulate matter, $PM_{10}$, $NO_2$, $SO_2$ and so on. At the same time, the degree vocabulary of the adjective needs to be established. Then match the data to it by using regular expression.

The list of high frequency words in EIA report is shown in the Table 1.

| No. | Evaluation indicator | No. | Degree description |
|-----|----------------------|-----|-------------------|
| 1   | $AQI_3$             | 1   | Great             |
| 2   | $PM_{10}$           | 2   | Good              |
| 3   | $PM_{2.5}$          | 3   | Normal            |
| 4   | $NO_2$              | 4   | Worse             |
| 5   | $SO_2$              | 5   | Severe            |
| 6   | $CO$                | 6   | Slight            |
| …   | …                   | …   | …                 |

Table 1. The list of high frequency words in EIA report

In general, it takes at least 500 words and sentences to establish a corpus and the next work is to mark and process this corpus. The processing of corpus includes the characteristic of word marking, syntactic marking and so on. There are some available software tools in this processing.

| No. | Tools name            | Description                        |
|-----|-----------------------|------------------------------------|
| 1   | Text filter           | Translate to plain text            |
| 2   | Text classifier       | Text categorization                |
| 3   | Proofing tools        | Standardized corpus                |
| 4   | Meaning of words marking | Marking the meaning of words    |
| 5   | Syntactic analysis    | Syntactic analysis of sentence     |

Table 2. Common tools of corpus processing

2.2. Subjective and Objective Empowerment
In general, subjective and objective weighting are two methods for weighting indicators. Subjective weighting emphasizes the expert experience in the processing of weighting and the results are often
subjective. Objective weighting method can reflect the difference of the indicators. This method is a kind of scientific weighting method but it is easy to ignore the results of subjective expectations. Therefore, this paper intends to combine the subjective and objective weighting method to achieve the purpose of selecting matching indicators.

The subjective weight vector and the objective weight vector should be considered comprehensibly. The attribute weight vector \( w_1, w_2, \ldots, w_n \) is obtained by the objective weighting method. \( 0 \leq w_j \leq 1 \) and \( \sum_{j=1}^{n} w_j = 1 \). The other attribute weight vector \( w'_1, w'_2, \ldots, w'_n \) obtained by the subjective weighting method. \( 0 \leq w'_j \leq 1 \) and \( \sum_{j=1}^{n} w'_j = 1 \). The \( m \) and \( n \) denote the importance of \( w \) and \( w' \). The total weight is \( w = mw + mw' \). It is the basic principle of subjective and objective weighting method. It fully considers the impact of subjective weighting and objective weighting on final results and makes the results more reasonable.

2.3. Regular Expression

The regular expression is the preferred tool for text preprocessing by NLP method. The text processing includes abstraction of stem and text, segmentation, text filtering and pattern matching. The purpose of regular expression is to describe a feature with a string. The string verifies other features and the strings. The text and data are regarded as ordinary characters. In the regular expression matching process, the matched character by subexpression is saved to the final matching result. The regular expression matching rules are from left to right. The next matching character should be performed after a character is matched. Normally, the control is taken from the expression and the position is matched from the string. The regular expression is considered as the suitable matching methods.

There are two matching engines in the regular expressions. One is the NFA (Nondeterministic Finite Automata) matching engine and the other is the DFA (deterministic Finite Automata) matching engine. Due to the fast matching speed, the most common method is DFA. Besides, NFA requires multiple backtracking during the matching process and the matching time is slower. Since the EIA report requires a high level of accuracy in the indicators, the DFA matching engine matches the text and data. It attempts to achieve high precision ambiguity detection. The flow charts of NFA and DFA are shown in the Figure 1 and Figure 2 [11].

![Figure 1. The flow chart of NFA](image1)

![Figure 2. The flow chart of DFA](image2)

From the figures, it can be clearly understood that the cost of constructing DFA is far greater than that of NFA. It is certain that constructing DFA will consume more time and memory. The NFA can support advanced features such as capturing group, look around and so on. But DFA cannot implement these functions.

3. Experiments

This section describes the specific sources of data and screens out the data that are useful for this article. The paper establishes a corpus that matches the text description and data in the EIA report and compares it with some common corpus. Compared with other Chinese corpora, the corpus provided in this paper has more advantages in matching speed and accuracy. After the corpus is obtained, the matching target is filtered. In the process of screening key index, this paper uses a combination of subjective and objective weighting methods. The subjective and objective weighting method shows more scientifically and objectively in selecting key index. After obtaining key index, it’s necessary to match text description
and data in the EIA report. The matching process uses the RE2 method of regular matching, which greatly improves the efficiency.

3.1. The introduction of Data
The data in this section is from the EIA report released in Haidian District, Beijing. The construction project is a real estate development project. The project was sourced from a 2015-year construction and development planning project in Haidian District, Beijing. The District covers an area of nearly 20,000 square meters. The EIA accounts for 0.15% of the total project investment. The EIA report contains information on topography, geomorphology, geology, climate, weather, hydrology, vegetation, etc. The paper selects the air environment status of the project site as the main research target and original input. The Table 3 is about the 2015 Air Quality Table in Haidian District of Beijing. The Table 4 shows the air quality summary.

### Table 3. 2015 Air Quality in Haidian District of Beijing

| Project     | $PM_{2.5}$ | $PM_{10}$ | $SO_2$ | $NO_2$ |
|-------------|------------|-----------|--------|--------|
| Annual average concentration ($\mu g / m^3$) | 80 | 102.9 | 15.2 | 56.1 |
| Standard value ($\mu g / m^3$) | 35 | 70 | 60 | 40 |
| Standard status | Excessive | Excessive | Reach the standard | Excessive |

### Table 4. Summary Sheet of Atmospheric Indices

| Monitoring project | Monitoring range | Standard Value (mg / m³) | Excessive rate | Maximum excess value |
|--------------------|------------------|--------------------------|----------------|---------------------|
| $PM_{2.5}$ (24h mean value) | 0.06-0.171 | 0.05-0.174 | 0.075 | 57.1 | 57.1 | 1.28 | 1.32 |
| $PM_{10}$ (24h mean value) | 0.101-0.266 | 0.07-0.225 | 0.15 | 57.1 | 14.3 | 0.77 | 0.5 |
| $SO_2$ 24h average | 0.023-0.048 | 0.016-0.049 | 0.08 | 0 | 0 | 0 | 0 |
| 1h average | 0.014-0.062 | 0.016-0.065 | 0.2 | 0 | 0 | 0 | 0 |
| $NO_2$ 24h average | 0.021-0.053 | 0.021-0.053 | 0.15 | 0 | 0 | 0 | 0 |
| 1h average | 0.010-0.060 | 0.017-0.060 | 0.50 | 0 | 0 | 0 | 0 |
| $CO$ 24h average | 0.8-1.4 | 0.8-1.4 | 4 | 0 | 0 | 0 | 0 |
| 1h average | 0.7-1.7 | 0.7-1.7 | 10 | 0 | 0 | 0 | 0 |
| $O_3$ Maximum daily average of 8h | 0.08-0.237 | 0.08-0.237 | 0.16 | 42.9 | 57.1 | 0.48 | 0.48 |
| 1h average | 0.069-0.338 | 0.08-0.336 | 0.2 | 35.7 | 35.7 | 0.69 | 0.68 |

The above table is a detailed collection method of air quality indicators, and two collection points are placed at the same time to collect data.

3.2. The Corpus of EIA
There are three kinds of corpus. They are Chinese corpus, multilingual corpus and Chinese general balance corpus. Their performances are shown in the Table 5.
Table 5. The comparison of corpus

|                        | Chinese corpus | Chinese-English corpus | Self-built corpus |
|------------------------|----------------|------------------------|------------------|
| Text description       | Support        | Support                | Support          |
| Data index of atmosphere (speed of matching) | slower         | slower                 | fast             |

It’s difficult for the common researches to provide a comprehensive, rich, timely, authoritative corpora material. Therefore, a special corpus should be established to achieve the above objectives. Since self-built corpus is more targeted, the matching speed is faster than others.

3.3. The experiments of Subjective and Objective Empowerment

This chapter compares the advantage and shortcoming (advantages and shortcomings) of subjective and objective weighting. For this paper, the single weighting method can’t meet the demand. The criterion of weighting results is the high frequency index in EIA. The weighting result is closer to high frequency index. It can be considered that this result is more effective.

Table 6. Comparison of several commonly used subjective and objective methods

| Advantage | Subjective | Objective |
|-----------|------------|-----------|
| AHP       | Simple and pragmatic | Has solid theoretical foundation |
| Delphi    | Combination of expert opinion | Too sensitive to some indicators |
| DEA       | The result influenced easily by subjective | Easy to ignore the intention of the subjective |
| EWM       | The credibility is not high because of less data | |

3.4. The experiments of Regular Expression

In this chapter, the traditional matching method and the regular expression matching are compared. Many cases will be considered by using traditional matching method and the way to match text description and data is complex. Instead, once set up the matching rule in regular expression, the instruction is automatically executed and the matching efficiency is greatly improved.
Table 8. The comparison of regular expression matching and traditional matching method

|                        | Regular Expression Matching | Traditional Matching Method |
|------------------------|-----------------------------|-----------------------------|
| Matching speed         | fast                        | slow                        |
| Code amount            | less                        | more                        |

Regular expressions are easier to use in text description and data matching than traditional matching methods because of the efficient matching engines.

4. Conclusion
This paper used some techniques in NLP to match the text description and data in EIA report and to achieve the purpose of ambiguity detection. In the selection of matching indices, the method of subjective and objective weight and the self-built corpus are selected, and then the method of regular expression can be used to match. This paper compared with the traditional method and got some good results. However, there are still some shortcomings in implementation, such as choosing a suitable regularization matching engine and the selection of the method of subjective and objective weights is still worth exploring.

Acknowledgments
This work was financially supported by the National Key R&D Program of China (No. 2017YFC0209901).

References
[1] Niesen T, Dadashnia S, Fettke P, et al. A vector space approach to process model matching using insights from natural language processing [J]. Multikonferenz Wirtschaftsinformatik (MKWI), 2016: 93-104.
[2] Nesi P, Pantaleo G, Tenti M. Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering [J]. Engineering Applications of Artificial Intelligence, 2016, 51: 202-211.
[3] Fredriksson K. On-line approximate string matching in natural language [J]. Fundamenta Informaticae, 2006, 72(4): 453-466.
[4] Sung S F, Chen K, Wu D P, et al. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study [J]. International journal of medical informatics, 2018, 112: 149-157.
[5] Badal V D, Kundrotas P J, Vakser I A. Natural language processing in text mining for structural modeling of protein complexes [J]. BMC bioinformatics, 2018, 19(1): 84.
[6] Fong A, Harriott N, Walters D M, et al. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events [J]. International journal of medical informatics, 2017, 104: 120-125.
[7] Feller D J, Zucker J, Yin M T, et al. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment[J]. JAIDS Journal of Acquired Immune Deficiency Syndromes, 2018, 77(2): 160-166.
[8] Sevenster M, Bozeman J, Cowhy A, et al. A natural language processing pipeline for pairing measurements uniquely across free-text CT reports [J]. Journal of biomedical informatics, 2015, 53: 36-48.
[9] Indukuri K V, Ambekar A A, Sureka A. Similarity analysis of patent claims using natural language processing techniques[C]/Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on. IEEE, 2007, 4: 169-175.
[10] Souli A, Cavallucci D, Rousselet F. Natural Language Processing (NLP)–A Solution for Knowledge Extraction from Patent Unstructured Data [J]. Procedia engineering, 2015, 131: 635-643.
[11] Aho A V, Lam M S, Sethi R, et al. Compilers: Principles, Techniques, and Tools (2nd Edition) [M]. Addison-Wesley Longman Publishing Co. Inc. 2006:74-84.