Research on the Importance of Data Enhancement Technology in Power Document Understanding

Ming Gao1, Jiayan Wang 2, Wenfei Zhang 3*, Dehui Wang 4, Zheng Peng 5, Hongzhi Lu6

123456Guangzhou Power Supply Bureau, Guangdong Power Grid Co., Ltd. Guangzhou, Guangdong, China, 510630

*1018205008@tju.edu.cn

Abstract. With the rise and development of the Internet and the artificial intelligence boom, natural language processing has been greatly developed in various fields and industries, including power industry. Intelligent document understanding, as a sub field of natural language understanding, uses artificial intelligence technology to enable machines to have natural language understanding ability. It has always been the focus of researchers and industry, and it is also the core problem of intelligent semantic interaction. As a common method to expand the sample set, data enhancement technology is an important text and image processing technology. Its core purpose is to use limited data and produce value equivalent to that from much more data. This technology has been widely used in various fields of deep learning. This paper reviews the development of data enhancement technology, lists and discusses five kinds of text enhancement technology paths and their corresponding representative technologies, including back translation, random word replacement, non-core word replacement, text enhancement based on context information, and text enhancement technology based on generative language model. It also analyzes the effectiveness of text enhancement technology in the aspects of regularization, transfer learning, improving model robustness, manifold and so on, to achieve the goal of improving the accuracy and effectiveness of document understanding in power industry.

1. Introduction

A great deal of data resources have been accumulated in the power field, including relevant standards and specifications, product and technical documents, management documents, fault resolution records, etc. How to search these documents quickly and intelligently is of great value to the operation, maintenance and fault recovery of power equipment. As far as language modeling tasks are concerned, common data enhancement methods include methods based on external data and random sampling methods based on Recur-Rent Neural Network LM [1-2]. Literature [3] points out that the foundation of realizing multi-objective optimization control of smart grid is the establishment of basic measurement and dynamic monitoring system, so smooth power system monitoring data is the key technology of establishing wide-area information sharing and wide-area measurement protection [4-5], and it is also the fundamental guarantee of realizing smart grid strategy.

The information retrieval system based on keywords has mature architecture design, sufficient theoretical basis, easy understanding in retrieval logic and mode and intuitive acceptance by users. In this paper, the discrete sequence generation problem is expressed as a reinforcement learning problem [6], and the generation model is regarded as a random parameterization strategy, and the output of the
discriminant model is used as a reward to optimize it, thus avoiding the problem that it is difficult to return the error gradient between the discriminant model and the generation model. By analyzing the importance of data enhancement technology in power document understanding, this paper studies and summarizes its influence rules, with a view to providing effective guidance for improving power system acquisition process and data processing methods.

2. A data enhancement algorithm based on RNN LM

Data enhancement technology is mainly used to improve the visual effect of images and improve the clarity of images. Or transform the image into a form more suitable for human or machine analysis and processing, so as to extract useful information from the image. The tourism domain knowledge base is constructed based on ontology technology, and the implicit and ambiguous information between concepts can be discovered through logical reasoning, and the implicit semantic association within domain resources can be discovered to realize intelligent search [7]. In this process, semantic matching is not carried out, and whether the query keywords submitted by users are semantically related to the results is not considered. On the basis of current reports and statistical analysis, it provides deeper semantic analysis and intelligent analysis query, explores the relationship between data, and establishes the semantic ontology model of data, which can better assist users' daily work and support power grid dispatching operation business.

The goal of RNN LM is to predict the conditional probability of each word in a given word sequence. Given a training sentence \( w_1, w_2, \ldots, w_T (w_t \in V, t = 1, 2, \ldots, T) \), \( V \) represents the dictionary space. RNN LM encodes the input word sequence as the hidden layer state sequence according to the following formula:

\[
 s_1, s_2, \ldots, s_T (s_t \in R^h, t = 1, 2, \ldots, T) \\
 s_t = \sigma(W_t s_{t-1} + W_v x_t + b_h) 
\]

In which \( x_t \in R^{|V|} \) is the unique hot code corresponding to \( w_t \), \( W_t \in R^{|V|h}, W_v \in R^{h|v}, b_h \in R^h \) is the trainable parameter, \( \sigma \) represents the nonlinear activation function, and then a series of conditional probability distributions are obtained by using the hidden layer state sequence

\[
 y_t = \varphi(W_o s_t + b_o) 
\]

where \( W_o \in R^{|v|h}, b_o \in R^{|v|} \) is a trainable parameter, \( y_t \in R^{|v|} \) represents the probability distribution \( P(w_{t+1}|w \leq t) \) of the current word \( w_{t+1} \) under the given historical sequence \( w_1, w_2, \ldots, w_t \), and \( e \) represents the soft-max activation function. MLE algorithm is usually used to estimate RNN LM parameters, that is, to maximize the logarithmic probability of a given training sequence

\[
 \sum_{t=0}^{T-1} \log P(w_{t+1}|w \leq t) 
\]

The storage of knowledge retrieval follows such an implementation process that the system classifies the input resources in terms of knowledge (concepts) instead of keywords, and then analyzes them in semantic level through natural language processing technology, so as to obtain the related concepts and knowledge categories of the resources, and finally stores them in the knowledge base for retrieval. Calculating the instability of signal is to strengthen the edge of ultrasonic signal from a statistical point of view. Its most prominent advantage is its low sensitivity to noise and high accuracy of edge detection.

3. Path of text enhancement technology and its corresponding representative technology

3.1. Back translation

The common practice of back translation practice is to translate a foreign language into the translator’s native language first, and then translate it backwards. Feedback information can regulate and control the
behavior of actors. Without feedback information, learning will be terminated and the original connection will be degraded. Facts have proved that the effect of timely feedback is better than that of delayed feedback. Such as self-study ability, understanding ability, abstract thinking ability, practical ability, observation ability, ability to analyze problems and solve problems. In the field of power dispatching, it is more necessary to retrieve structured data, especially around the power grid equipment, and search the model data of the power grid equipment, such as equipment operation data, management data and so on. If a keyword appears many times in a document, that is, it appears frequently, it can also be considered that the document has a high degree of correlation with keywords. In recent years, scholars at home and abroad have designed a series of new methods to calculate the correlation degree, which can improve the retrieval accuracy in the retrieval system.

3.2. Random word substitution
In the process of text classification, the dimension of text feature vector is too large, which leads to large storage space and slow processing speed, which will bring great overhead to training and classification time. Secondly, in fact, a large part of these words have nothing to do with categories and have little effect on classification.

If $x = \{0,1\}^n$ is the feature space, $\gamma$ is the category. Given a word sequence, $x_i = \{a_1, a_2, \ldots, a_n\} \in x(i = 1,2,\cdots,t)$, which represents the feature vector of a document, $\gamma \in \gamma$ represents a specific category. Let $S = \{0,1\}^m$ represent another feature space, given a keyword sequence, let $s_i = \{\delta_1, \delta_2, \ldots, \delta_m\} \in S(i = 1,2,\cdots,t)$ represent the feature vector value of the keyword.

Random word replacement technology can include some keyword information into the training text feature space, and effectively use the keyword information in the text classification process to improve the performance of the classification system [8]. If the proper partial order relation is defined on the set of part-of-speech markers, then the part-of-speech marker string (hereinafter referred to as marker string) is a sequence with order characteristics. We can use random word replacement technology to reduce the dimension of text feature space, and map the text to a set of conditional probability vectors of keywords. The new text feature space can well cover the information of text keywords.

3.3. Non-core word substitution
In view of the acquisition of polysemous words with high frequency, there is a view that paying attention to non-core word replacement is an effective strategy in the cognitive process of polysemous words, which can enable learners to guess the diffraction meaning of polysemous words more accurately and quickly according to the context, and can also strengthen learners’ long-term memory of the diffraction meaning, thus improving the effect of incidental vocabulary acquisition. Emphasis does not require a component to be graded when enhancing its meaning (but not its degree). However, when the emphasized components can be graded, adverbials play the role of reinforcement. Incidental learning and explicit learning are complementary to each other. Especially for the acquisition of non-core word replacement of high-frequency words, the combination of the two learning methods is more conducive to promoting the development of vocabulary knowledge of high-frequency words. These words are mainly used to reinforce adjectives and verbs. For the convenience of the following description, the reinforced adjectives are divided into two categories: A adjectives with absolute meaning (including extreme words and words with negative meaning); B is a large number of other graded adjectives.

3.4. Text enhancement based on context information
In Web information retrieval, there are usually two different retrieval tasks: known web page location and unknown information query, while the task of “unknown information query” is that users want to find information about a certain topic, but they are not sure whether this information exists or not. Users are concerned about whether the content of the web page is related to the information they want to query. Since such a context is not a natural context, we call it a special context. For example, through Google
Academic, DBLP, Pub Med, etc., it can match any information in the abstract, but IR can't solve the problems of identifying information types, replacing synonyms of keywords, and defining hierarchical relationships among keywords.

There is no doubt that text enhancement based on context information can be regarded as a special problem of text classification, which can distinguish whether the views expressed in natural language of text are positive positive appreciation or negative negative criticism. Usually, the link text of a webpage can express the theme of the webpage centrally and accurately, and even many link texts are the URL address of the webpage, which is consistent with the requirements of the webpage positioning task.

It is assumed that the short text is described by the attribute value \( \{w_i\}_{i=1}^m \), and the target attribute set \( C = \{c_1, c_2\} \), in which \( c_1 \) is in favor and \( c_2 \) is against. The classification decision for the short text to be classified is [9]:

\[
H_{NB} = \arg\max_{c \in C} \prod_{i=1}^n P(w_i|c)
\]

Among them, the calculation of \( P(w_i|c) \) adopts the Multinomial model. By correctly estimating \( P(c) \) and \( P(w_i|c) \) values, the emotional polarity of short texts to be classified is discriminated.

IE text mining benefits from the structured representation of information, which can provide semantic extraction syntax and transform unstructured information into structured information with the help of templates or preprocessing operations. However, the extraction relies on file scanning, and its efficiency is low. Theoretical analysis shows that under certain conditions, the text enhancement method based on context information based on domain attribute division can reduce both sampling error and approximation error [10].

### 3.5. Text enhancement technology based on generative language model

Machine learning generally refers to the process of transforming one language into another by using computers. Machine learning has been studied for more than half a century. Soon after the invention of computer, many scientists and inventors successively put forward the theoretical and practical plans or ideas of machine learning. The formation process of language concept is usually from fuzzy, global semantics to precise, local semantics, thus determining the complete semantics. The basic idea of statistical machine learning method based on source-channel model is to regard machine learning as a process of information transmission, and the information passing through the channel means that the information has been translated. Because the optimization is carried out at the semantic level, the conditional body mainly judges the same or different semantic information of predicates, parameters and modifiers of two adjacent predicate expressions, such as semantic classification code and semantic roles.

The results show that the increase of language model training corpus can steadily improve the effect of machine translation. However, the position and role of these semantic components in sentences and the corresponding vocabulary and other linguistic information in the generated texts in different languages are still undecided. The basic characteristic of this process is that it is a function of time, but the value observed at any moment is uncertain and a random variable. Grammar resource mapping is to determine the position of semantic components that identify sentence content in sentences according to the characteristics of grammar resources in different languages. However, the construction of semantic language model is much more complicated, and it does not show better performance than N meta-language model in many natural language processing tasks, so N meta-language model is the most widely used statistical language model at present.
4. Application of data enhancement technology in power document understanding

4.1. Syntactic rules based on complex feature sets

The position of each word in different sentence patterns may be different. According to the fitting degree between the application system and the model, the advantages and disadvantages of the language model are evaluated. Firstly, the short text to be tested is classified into suitable fields; Secondly, the corresponding intra-domain emotion discriminator is used to classify and judge, and the judgment result is output. The embedded patterns such as PAT array and prefix-suffix index structure store the text content and position information together, and BWT algorithm is used to compress the index to reduce the storage space. This kind of method has poor dynamics and is not ideal for dealing with long text indexes.

An important task of the query analysis module is to expand the search keywords submitted by users based on ontology. The method used in this paper is to expand the semantics through ontology knowledge base, and the core algorithm is the semantic similarity calculation model. In order to fully describe the implementation process of semantic expansion of search conditions in ontology-based smart grid knowledge retrieval system, the related processing flow of semantic expansion involved in this paper is described as follows. Figure 1 shows the main flow of this semantic extension.

![Figure 1 Ontology-based semantic extension process](image)

The user response module accepts the input of the user's search request and sends the user's search request to the query analysis module. The query analysis module needs to establish an ontology-based semantic similarity calculation model in advance and set a reasonable semantic similarity threshold, so as to ensure that the extended keyword set is screened by semantic similarity.

The complex feature set used in the system is the set of the following pairs:

\[ A = \{ (F_1, V_1), (F_2, V_2), \ldots, (F_n, V_n) \} \]  \tag{5} 

In which \( A \) is a complex feature set and \( (F_i, V_i) \) is the \( i \) th \( \{ \text{Characteristics, Values} \} \) pair, which indicates that the attribute feature \( F_i \) has an eigenvalue \( V_i (i = 1, 2, \ldots, n) \). "and" and "," They represent the AND and OR relationships between the elements in the set. The eigenvalues of a feature \( F_i \) in complex feature set \( A \) can be found by path search.

The module of obtaining sentences refers to how to segment a sentence. Sentences can be divided into three levels, in which words constitute the smallest unit of sentences, words constitute phrases, and the next level of phrases is the smallest translation unit. For example, scheduling management class,
power grid operation class, equipment model class, etc., then the classified index files are merged, and
Lucene can improve the index efficiency in the process of creating index by using the hardware
resources of the machine. At the same time, the influence of prefix subsequence in the generated
sequence on the generation of subsequent entries is also considered, and the unstructured information of
the extracted objects is segmented to identify named entities. Using the knowledge model and the
context relationship of entities, the types of entities are identified. Entities are extracted and then stored
in the entity library.

4.2. Performance enhancement method based on domain attribution division
The text emotion polarity classification method based on domain attribution division is modeled as a
two-stage processing process: the first stage (inter-domain attribution classification process):
determining the domain to which the short text to be processed belongs; The second stage (intra-domain
emotion discrimination process): emotion discrimination is performed on the short text to be processed
by the classifier in the corresponding domain. Word features and position information are stored
separately in the separation mode. For example, the posting file with inverted index records the position
information in the text where the word is located separately, which takes up about 3.0 ~ 4.0 times of
dictionary information. The traditional method of filtering and querying through each functional module
is not only inefficient in querying, but also requires a lot of labor time. Therefore, the intelligent retrieval
technology of power grid dispatching knowledge base can improve the retrieval efficiency and accuracy
of power grid dispatching management system.

In the test, semantic search will be converted into SPARQL language of RDF for query, and keyword
search will be converted into SQL statements containing Like keywords for query. Full-text index
search of Like keywords can be performed by using a database supporting full-text search (see Figure 2
and Figure 3).

![Figure 2 Precision ratio](image-url)
Figure 3 Recall rate

On the whole, semantic search can meet the search needs of more users, and achieve higher precision. It can combine fields to customize and analyze complex query conditions.

In the course of confrontation training, the selection of superparameters is very important to the performance of the finally generated model. In this process, due to clock jitter, phase-locked error and signal transmission delay, the random disturbance of synchronous sampling signal is caused, which makes the sampled data contain noise caused by phase jitter, and the phase noise is different from white noise. Because all personnel information can be semantically analyzed in advance and semantic tags can be added, this type of semantic search can be supported. This type of keyword search has no return results, so the search rate is 100%.

5. Summary
The safe operation of power grid system needs the important guarantee of power facilities. If the power equipment is damaged, it will bring trouble to people's lives, and even affect the economic development and social stability and harmony. Different from the traditional generation task, in the language model enhancement task, it is necessary to ensure the diversity of generated text data, even if the generated model maintains a certain generalization performance. Data enhancement technology can not only effectively improve the generalization ability of neural machine translation for languages with poor resources, but also improve the translation quality for languages with rich corpus. Power document understanding is a complex project, and there are still many technical challenges in each link. This paper only gives some preliminary ideas and methods. If we want to achieve efficient and accurate automatic knowledge extraction, we need to combine natural language understanding, deep learning and other related technologies.

References
[1] Zhang Xiaofeng, WU GANG. Data Enhancement Method Based on Generative Confrontation Network [J]. Computer System Application, 2019, v.28(10):203-208.
[2] Liu Yufei, Kelly Y Zhou, Liu Xin, et al. A new generation of artificial intelligence data enhancement method based on Wasserstein GAN —— Taking cancer staging data in biological field as an example [J]. Engineering, 2019, V.5 (01): 338-354.
[3] Gu shuhao, shan Yong, Xie Wanying, et al. neural machine translation technology based on data enhancement and domain adaptation [J]. journal of Jiangxi normal university (natural science edition), 2019, 043(006):643-648.
[4] Long Ying, Zhang Enjia. Research prospect of intelligent planning under the framework of data enhancement design [J]. Urban Planning, 2019, v.43; No.392(08):36-42+54.
[5] Wu Tianyu, Xu Yingchao, Chao Pengfei. A data enhancement technique to improve the accuracy of target image recognition [J]. Journal of Laser, 2020, v.41; No.272(05):100-104.
[6] Wu Xue, Song Xiaoru, Gao Song, et al. Convolutional neural network fire identification based on data enhancement [J]. Science, Technology and Engineering, 2020, 020(003):1113-1117.

[7] Zhang Hao, Chen Long, Wei Zhiqiang. Abnormal traffic detection technology based on data enhancement and model updating [J]. Information Network Security, 2020, 000(002):66-74.

[8] Liu Li, Chen Yuhui, Liu Xiaoting. A technical research on recognition and classification of engineering drawing documents [J]. Electronic Design Engineering, 2020, v.28; No.434(12):95-100.

[9] Zhang Yike, Zhang Pengyuan, Yan Yonghong. Language model data enhancement technology based on confrontation training strategy [J]. Acta Automatica Sinica, 2018, 44(05):126-135.

[10] Gao Youwen, Zhou Benjun, Hu Xiaofei. Research on Convolutional Neural Network Image Recognition Based on Data Enhancement [J]. Computer Technology and Development, 2018, 28(008):62-65.