

# Study of Intelligent Search Engine of Energy Industry Based on BERT Semantic Model

Jiayang LI<sup>1</sup>, Hao LI, Ni YAN, Ziyun CHEN

*Nari Group Corporation (State Grid Electric Power Research Institute), Nanjing  
Jiangsu 210061, China*

**Abstract.** Based on a variety of heterogeneous data of energy enterprises, this paper provides an intelligent search method based on the Bert preprocessing model. This paper provides a method of intention recognition based on the combination of template matching and text classification, which involves machine learning, deep learning and other fields of artificial intelligence. First, the preprocessing model Bert is used to rewrite the natural language into a vector based on syntax. Then, the information extraction technology is used to extract the structured and machine understandable information from the problem to provide parameters for intention processing. The intention recognition technology used in this paper first uses template matching method, and then uses text classification method. Finally, the effects of various methods are compared through experimental examples.

**Keywords:** BERT model, word segmentation, search engine, deep learning

## 1. Preface

China's economic development has entered the "new normal", which is changing from high-speed growth to medium high-speed growth, and new changes have taken place in the energy industry. With the emergence and development of Internet of things, digitalization and online technology, the integrated dispatching construction of coal power energy enterprises has begun to transform to digitalization, networking and intelligence. The business of energy enterprises is complex, and there are a large number of structured, semi-structured data in the production and involving power, chemistry, railway, port, aerospace, petroleum and other professional fields. The data of energy enterprises has a large amount of data and complex structure, which is one of the difficulties in data analysis and retrieval.

Google team released a large-scale pre training language model [1] (Bidirectional Encoder Representation from Transformers, BERT) based on bidirectional transformer at the end of 2018 and this model refreshed the current optimal performance record of 11 NLP tasks by virtue of the pre trained fine tuning model. The data of energy enterprises records a large number of business charact. At the same time, the text data also contains a

---

<sup>1</sup>Corresponding author: Jiayang Li, Nari Group Corporation (State Grid Electric Power Research Institute), Nanjing Jiangsu 210061, China; E-mail: 429901895@qq.com.

lot of industry terms, involving a wide range of fields, and has a huge amount. So, we need an effective intelligent search method to quickly retrieve the useful text data and find out the effective data information in the massive data.

According to the characteristics of the data in the energy industry, this paper focuses on the key technologies and core algorithms in the application of natural language processing, uses the pre training Bert model to analyze the content of questions and text data, mines the context rules of text data, and conducts secondary training through a large number of Chinese pre training models. The goal of this paper is to use natural language processing technology to transform questions suitable for the energy industry into machine language, provide reference suggestions for the application of natural language pre training model in the energy industry.

2. Model introduction

In this paper, the structure of the model is fine tuned based on the pre trained Bert line, so that it can be better applied to natural language processing and text classification tasks. We need to have a more comprehensive understanding of the structure and characteristics of the Bert model in order to make better use of the preprocessing model. From the perspective of professional technology, the Bert model does not use a new deep network or propose a new training algorithm, but improves and integrates the existing deep learning algorithm and natural language processing technology. Next, we introduce the Bert model.

The full name of Bert model is Bidirectional Encoder Representation from Transformers. As shown in Figure 1, the Bert model [2] is mainly divided into two parts, the input layer and the bidirectional transformer coding layer. When processing a word, we can also consider the words before and after the word to get its meaning in the context. The attention mechanism of transformer has a good effect in feature extraction of words in context, and intuitively, considering the bidirectional encoding of context is better than only considering the unidirectional effect of the above (or below).

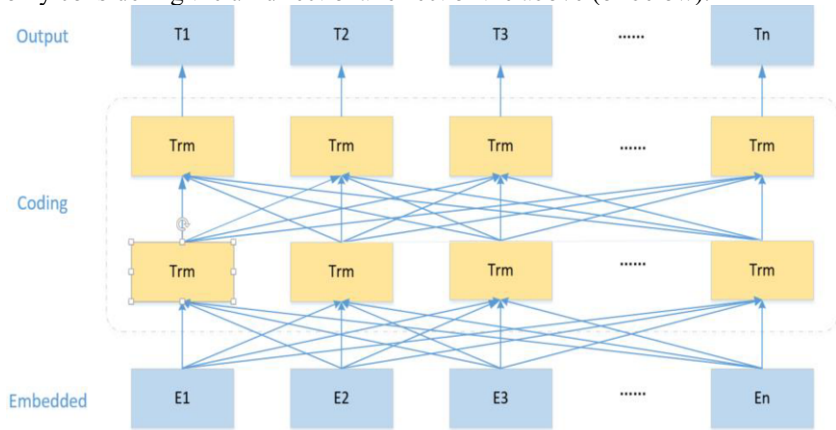
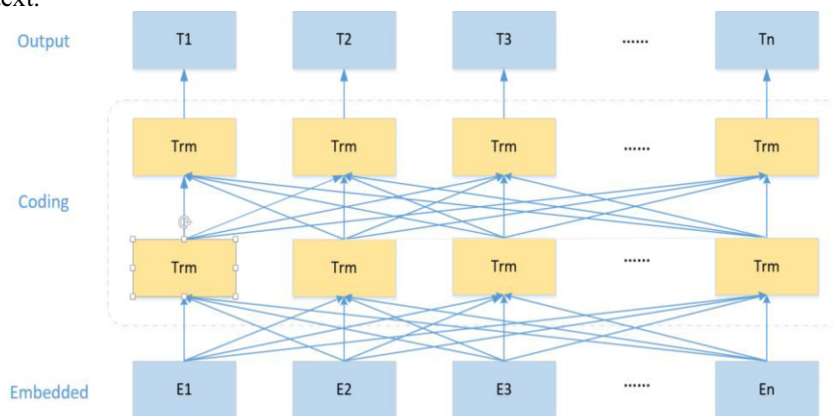


Figure 1. Basic structure of BERT model.

In natural language processing, word embedding is the basis of all researches. Using a vector to represent words and capture the semantics and relationships of words. A sentence is not a simple combination of some words, it must conform to human language

habits, and the language model is used to determine whether a sentence conforms to human habits, specifically, the word in front of a word (or context) in a given sentence, and then predict the probability of the word. Take an example of Word 2vec, which is proposed by Tomáš Mikolov in 2013, its feature is that all words are represented as low dimensional dense vectors, which can qualitatively measure the similarity between words in the word vector space. The core idea of word2vec is to hide a word from a sentence and train neural network to predict the hidden word. When the neural network training can successfully predict after a large number of corpus training, the word vector representation of the word can be settled through the neural network. However, with the increasing complexity of natural language processing tasks, word2vec can only return a word vector mixed with multiple semantics, which cannot solve the problem of polysemy. So, based on the problem that word2vec model cannot solve, researchers propose a new word embedding method (Embedding from Language Models, ELMO). In Elmo, the pre-trained model is no longer just a vector correspondence, but a trained model. When using Elmo model, input a sentence or a paragraph into the model, and the model will infer the corresponding word vector of each word according to the line text. One of the obvious advantages of this is that we can understand polysemous words in context.



**Figure 2.** Basic structure of Elmo model.

As shown in Figure 2 [3], each word in the sentence is encoded in three levels: In the coding layer, the first layer is the corresponding word coding in the forward LSTM, which mainly contains syntactic information; then there is the second layer of reverse LSTM, which contains semantic information. Finally, the embedded layer and the two coding layers are given a weight respectively, and the coding results are multiplied by their respective weights and summated to get the output of Elmo.

Combined with Figure 1 and Figure 2, it can be seen that there is no essential difference in the overall structure between Bert model and Elmo model. The fundamental difference is that the best model uses transformer encoder instead of LSTM encoding of Elmo model. The prototype of transformer includes two independent mechanisms, one encoder is responsible for receiving text as input, and the other is responsible for predicting the results of the task. BERT model is a multi-layer transformer encoder based on fine tuning, whose goal is to generate language models, so only the encoder part of transformer is required. By using transformer instead of Bi LSTM as encoder, Bert can have deeper layers and better parallelism.

3. Construction of Intelligent Search Engine Model

3.1. Transformer Encoder Based on Bert Model

Calculation steps of self-attention output:

- 1) For each word vector of the input encoder, multiply by three parameter matrices to create three vectors (query vector, key vector and value vector).
- 2) Calculate attention score: calculate the degree of attention a word places on words in other positions in the sentence, and calculate it through the dot product of the query vector corresponding to the word and the key vector of each word.
- 3) Divide each score by  $\sqrt{d_{key}}$ ,  $d_{key}$  is the length of the key vector.
- 4) These scores are normalized by a softmax.
- 5) Multiply the normalized score by each value vector.
- 6) Add the values in the previous step to obtain the output in the self attention layer.

In this paper, a fully connected network is added to the output of Bert and softmax is used as a classifier [4]. According to steps 2) - 6) above, the output of self attention can be obtained. As shown in Figure 3, the left part is the encoder, which is used to encode the source sequence; the right part is decoder, which is used to decode the output target sequence. As can be seen from the encoder on the left side of Figure 3, it is composed of N layers stacked together, and the structure of each layer is the same. Each layer consists of two sub layers, the first sub layer is composed of multi head attention sub layer and feed forward neural network sub layer. The second branch consists of a simple fully connected feedforward network.

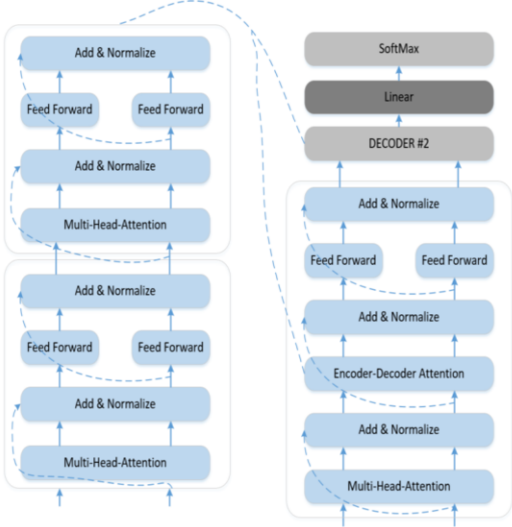
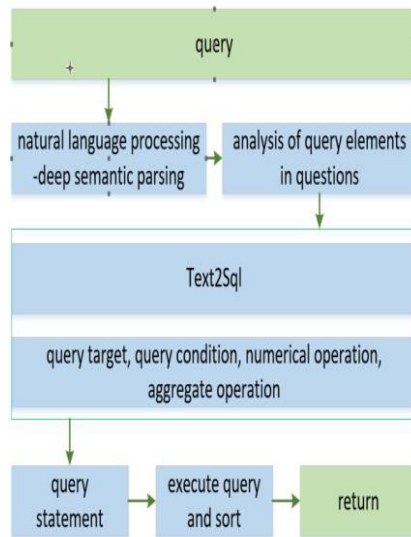


Figure 3. Transformer structure.

The decoder on the right is also made up of N-identical-layers. In addition to the two branch layers of the encoder, each layer adds a third branch layer coding decoding attention sub layer [5] (Encode-Decoder Attention), it is used to calculate the influence of output results in encoder on current results. In addition, the first multi head

self-attention sub layer is not exactly the same as the sub layer of encoder. Mask structure is added to the encoder to deal with the words shielded in the training stage. The decoder also performs residual connection and layer normalization.

### 3.2. Intelligent Search Engine Architecture



**Figure 4.** Search engine process.

The processing flow of search engine is described. As shown in Figure 4, according to the user's question, the intelligent search engine uses natural language processing technology to analyze the element information in the question, recognize and classify the text [6]. The natural language is transformed into SQL statements, and the returned results are sorted. Current search engines tend to pay attention to generality, but invest less in the acquisition and analysis of user query requests. Usually search engine only supports keyword search and logical operation on this basis [7]. However, the search engine proposed in this paper has made the following two improvements in data preprocessing:

(1) Using natural language processing technology, Extract the effective components of query conditions, including lexical information and its logical relationship, to prepare for intention recognition.

(2) Through the study of superior words and inferior words, according to the user-defined dictionary to get synonyms, synonyms and Related words of keywords.

The above analysis is mainly for vocabulary. When you enter a question with complex logical relationship, it is difficult to combine the search terms with keywords. So, the use of natural language processing technology is the ultimate and fundamental way to achieve the accurate description of the search and ensure the search relevance and similarity. Next, let's introduce the detailed steps of intelligent search engine in this paper:

1) Data input: A user's question is described as a piece of data, so that the data set to be trained is  $Zy = \{Zy_1, Zy_2, Zy_3, \dots, Zy_n\}$ ;

- 2) Chinese word segmentation: This paper uses the CRF word segmentation algorithm [8], at the same time, it updates the energy industry based terminology and dictionary to segment the question text  $A_{ij} = \{A_{11}, A_{12}, A_{13}, \dots, A_{yn}\}$ ; one question corresponds to the formula I, each of which is  $j$ ; Here, the conditional probability calculation formula represented by CRF model is:

$P(Y|x; \theta) = \frac{1}{Z_x(\theta)} \exp\{\sum_{i=1}^k \theta_i f_i(D_i)\}$  where, the denominator is the partition function, and the expression is:  $Z_x(\theta) = \sum_Y \exp\{\sum_{i=1}^k \theta_i f_i(D_i)\}$ , when using the optimization method to train the parameters of CRF model, the main task is to calculate the cost and grad expressions of the model.

- 3) Data preprocessing: Remove redundant words in the text, such as non-functional auxiliary words, adverbs, prepositions, conjunctions and so on.
- 4) Text representation: Using Bert pretreatment model, each word after text segmentation is transformed into a vectorized value, Using the vectorized value to find the optimal matching result;
- 5) Feature construction: construct the corresponding feature thesaurus matrix to complete feature extraction, which can effectively reduce the dimension of feature space and improve the operation efficiency:

$$\sum A_{IJ} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1j} \\ A_{21} & A_{22} & \dots & A_{2j} \\ A_{i1} & A_{i2} & \dots & A_{ij} \end{bmatrix}$$

6) Model training: in this method, three models need to be trained, that is, the classification model of predicting table names; The classification model of predicting column names and the model of converting to SQL statements. The feature matrix is used as the input of the algorithm model, of which 70% is the training set and 30% is the test set;

7) Evaluation process: This paper uses the accuracy, recall, precision and other four indicators to evaluate the results. In fact, the accuracy rate is for our prediction results, which indicates how many of the predicted positive samples are really positive samples. Then there are two possibilities for positive prediction. One is to predict the positive class as the positive class (TP), and the other is to predict the negative class as the positive class (FP). Recall rate indicates how many positive examples in the sample are predicted correctly. The accuracy reflects the closeness of the measurement result to the true value.

■accuracy:

$$ACC = \frac{TP+TN}{P+N} = \frac{\text{number of true predictions} + \text{number of false predictions}}{\text{number of correct predictions} + \text{number of errors}}$$

■recall:

$$R = \frac{TP}{TP+FN} = \frac{\text{Number of true prediction}}{\text{number of true prediction} + \text{number of false prediction errors}}$$

■precision:

$$P = \frac{TP}{TP+FP} = \frac{\text{Number of true prediction}}{\text{number of true prediction} + \text{number of false prediction}}$$

#### 4. Experimental Data and Analysis

In the production and operation practice of the national energy group, priority is given to ensuring the continuity of coal production and coal transportation process and maximizing the utilization of production capacity, so as to "promote" the efficient operation of the industrial chain. In the process of coal production and transportation, multi link manual coordination is carried out to dynamically balance the coal demand and ensure the coal supply of internal and external power plants. The overall industrial chain operation belongs to "production promotion", Production and operation mode of "two-way balance between supply and demand". The training data of this paper are the data of electric power, coal, integration, sales, chemistry, railway, port, shipping, oil and other modules in the energy industry. The data of each category are shown in Table 1.

**Table 1.** samples.

field	category	samples
coal	coal_t	12003
power	pow_t	11908
railway	rail_t	9872
port	port_t	8794
ship	ship_t	8976
chemical	che_t	9321
sale	sail_t	10023
integrated	uni_t	6532

The model training is carried out on Tesla V100 GPU, and the training parameters are shown in Table 2.

**Table 2.** training parameters.

parameter	category
network layers	24
Hidden layers	768
Self-Attention-Head	12
Parameters	340M
Activation function	GLUE
epochs	50

The learning and training of the algorithm model adopts the method of repeated sampling with putting back to construct the training set and the test set, in which the training set and the test set are divided according to the proportion of 70% and 30%. In the training process of this paper, in order to train a deep two-way language vector, we need to cover some words in the sentence. The shielding strategy is 80% probability to keep the word unchanged, 10% probability to replace with a random word, and 10% probability to replace with mask token.

Three models will be trained in this paper. The first model is the classification model to determine the table name; The second model is to determine the column name; The third model is the model file that generates SQL.

In order to reasonably evaluate the effect of intelligent search semantic analysis, we focus on the accuracy, recall and accuracy of the model. In addition, in order to compare the effect of BET model and traditional machine learning classification model, this paper selects several typical machine learning algorithms for comparative experiments. In this paper, LSTM long-term memory network and RNN recurrent neural network are used for experimental comparison. The results are shown in Table 3:

**Table 3.** experimental result.

Algorithm model	Accuracy	recall	precision;	time
Pre-training model BERT	0.9551	0.899	0.832	392s
LSTM	0.921	0.833	0.765	345s
RNN	0.889	0.863	0.835	492s

It can be seen from table 3 that the training time of LSTM using long-term and short-term memory network is the least, but the accuracy, recall and precision indicators are not as good as the pre training model Bert. The pre training model Bert has good performance in accuracy, recall, accuracy and other indicators. The pre training model Bert can bring the text features of domain training to the scene of search engine, and can significantly improve the accuracy.

## 5. Conclusion

In the intelligent search engine based on energy system data, by introducing the Bert model based on pre training, and comprehensively comparing the pre-processing model with the commonly used machine learning and deep learning models, by comprehensively comparing the accuracy, recall, accuracy and time-consuming of other algorithm models, and training the model suitable for energy system, an optimal scheme of intelligent search on national energy system is given.

At present, in addition to the more mature Bert model, there are a series of more complex preprocessing models such as mass and xlnet. It can be expected that with the development of artificial intelligence, natural language processing and other technologies, the intelligent search engine based on the national energy system business will usher in great changes.

## Reference

- [1] Hu CT, Qin JK, Chen JM, Zhang L. Research on application of public opinion classification based on Bert model. *Network Security Technology and Application*, 2019 Nov; 2019(11):41-44.
- [2] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language und erstanding [J] 2018.
- [3] Tomas M, Kai C, Greg C, and Jeffrey D. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.
- [4] Ren Y. Power grid a classification model of power work orders texts based on pre-trained BERT model. *Yunnan Electric Power*. 2020; 048(001):1-7,11.
- [5] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences [J]. *Eprint Arxiv*, 2014, 1.
- [6] Liu XH, Chen WS, Ren QN, News text classification based on attention cnlstm network model. *Computer Engineering*. 2019 Jun; 56(14):7.
- [7] Vaswani A, Shazeer N, Parmar N, et al. Attention I s All You Need [J]. 2017.
- [8] Zhang K. Application and development of transmissionline inspection system based on multi-rotor UAVs [J]. *Modern Industrial Economy & Informationization*, 2013.