Data Mining and Intelligent Matching Algorithm in Electric Power Terminology Database

Yanan Chen*, Yuxi Lin, Luebin Fang, Xiuyuan Cheng, Weilan Wang
State Grid Fujian Information & Telecommunication Company, Fuzhou 350001, China

*Corresponding author e-mail: yananchen@sgit.sgcc.com.cn

Abstract. The paper designs a method and system for automatic extraction of professional vocabulary in the electric power field based on data mining and matching algorithms. The system extracts professional vocabulary related to power marketing from the corpus in the field of power marketing, converts it into corresponding English and adds it to the professional vocabulary of power marketing. Then we will collect the relevant vocabulary of the power system and the corresponding translation and add it to the professional vocabulary of power marketing. The thesis establishes a two-level index structure for the professional vocabulary of electric power marketing. We search the professional vocabulary database of electric power marketing, the order of retrieval is electric power system vocabulary, general dictionary, electric power marketing field extraction vocabulary, and the search result is returned to the user after the retrieval. The experiment verified that the system facilitates the efficient retrieval of user vocabulary, and then realizes the automatic conversion of professional vocabulary in the marketing standardization design results, so as to improve the quality and efficiency of professional vocabulary conversion.

Keywords: Data mining matching algorithm, electric power terminology database, automatic extraction.

1. Introduction
Terminology is an agreed symbol that expresses or defines professional concepts through speech or text. It is the crystallization of human scientific knowledge in language and plays an important role in the development of science and technology, cultural heritage and social progress. In addition to the name function, semantic function, communicative function and pragmatic function, terminology also has cognitive function, tool function, intellectual function, accumulation function, legal function and information compression function. In the contemporary social life of globalization and informatization, the importance of terminology translation has become increasingly prominent, and related researches have also increased. However, there are still few special discussions on the construction and application of bilingual terminology databases in electric power. From the perspective of translation practice in the electric power industry, terminology databases, especially those in electronic form that can be directly called by CAT software, have become the key and scarce resources to improve industry translation efficiency and improve industry translation quality together with industry translation memory [1]. Based
on the actual needs of the electric power industry, this article introduces the role and construction process of this type of terminology database, and provides references for improving the translation efficiency and quality of electric power professional texts.

2. The idea of extracting professional terms
The traditional method of extracting terms from the Chinese-English parallel corpus is to first identify the respective noun phrases of Chinese and English, and then perform the corresponding method. Due to the different grammatical characteristics of Chinese and English and the inconsistency of the recognition standards of Chinese-English professional terms, many professions the term cannot correspond to or the semantically inconsistent situation thereafter.

There are two main problems with such methods: one is that most of them contain one-to-one hypotheses corresponding to words. But this is not the case [2]. The English-Chinese translation contains a large number of one-to-many and many-to-many word correspondences, such as "floppy drive/floppy drive", "system file/system file" and so on. Another problem is indirect correlation. Indirect correlation means that some bilingual words that are not directly corresponding have a high co-occurrence probability due to the influence of monolingual fixed collocations and compound words. For example: "opening/operation" and "system/system" are two correct translation equivalents. However, due to the strong collocation relationship between "operating system" and "operating system" in a single language, the translation pairs "system/operation" and "operating/system" also have a high co-occurrence probability and are extracted as translation equivalent pairs. Indirect correlation greatly reduces the correct rate of translation equivalent pair extraction.

In order to solve the impact of multi-word correspondence and word segmentation errors, here we use the N-gram model to extract candidate translation units. First, use the Java word segmentation system provided by the Chinese Academy of Sciences to perform Chinese word segmentation on the acquired corpus, use stop words to divide the sentence into chunks, and then perform N-gram extraction within the chunks, that is, each word and it's the combinations of N words adjacent to each other in the block are regarded as candidate translation units (the value of N can vary from 1 to the length of the block). Here stop words refer to those words that are weak in word formation and appear frequently.

After obtaining the candidate translation unit, the corresponding relationship is determined by counting the co-occurrence probability of the Chinese-English candidate translation unit in the bi-sentence pair. After calculating the translation probability of the candidate translation pair, the equivalent pair is extracted according to the probability. Finally, the greedy hypothesis is used to solve the indirect correlation problem [3]. This hypothesis considers that the source language and target language vocabulary in the same sentence pair may become translation equivalent only when there is no candidate equivalent pair with higher translation probability related to these two vocabularies. The structure diagram of the term extraction system is shown in Figure 1:
As can be seen from Figure 1, the system mainly includes the following modules: Text reprocessing module: This module uses the Java word segmentation system provided by the Chinese Academy of Sciences to segment the acquired corpus, use Chinese and English stop words to divide blocks and use N-Gram model generates bilingual candidate translation units. Extraction module: According to the algorithm idea of extracting professional terms described above, calculate the translation probabilities between candidate translation equivalent pairs according to three commonly used statistical co-occurrence functions (i.e., Dice coefficient, Phi square coefficient and LLR), and the translation probability is used to extract terms and compare the extraction results. Result database: store and manage the extraction results in the form of relational database organization.

3. Description of the problem of the electric power terminology database of the data mining matching algorithm

The data mining matching algorithm is a network form with cognitive computing as the core, and its essential feature is the network node and/or the global feedback control of the cognitive loop feedback control structure MDE, as shown in Figure 2. This paper proposes to use electric power the idea of professional terminology database is to transfer the electric power terminology database on network nodes that have random failures or resources that do not meet the conditions to other nodes. At the same time, the autonomy of data mining matching algorithms provides migration for failure services from one node to other nodes [4]. The calculation ability and state temporary storage ability of the path lay the foundation for the realization of the electric power terminology database.
Figure 2. Cognitive feedback control loop structure

The transferable service set \( M \) can be represented by a power terminology term frequency and power terminology term frequency, which is defined as \( M = (H, U) \), where \( H = \{H_i \mid i = 1, 2, \ldots, n\} \) represents a set of \( n \) services in the network. \( U_{ij} = \{(H_i, H_j) \mid H_i, H_j \in H, i < j\}, \mid U \mid = e \) represents a directed edge set with \( e \) edges. Each node in the term frequency of electric power terminology represents a service, which is the smallest unit in the terminology database of electric power. The weight of the term frequency service node \( H_i \) of electric power terminology is the calculation cost, which is recorded as \( \overline{W}_{ij} \), represents the time dependency between services. Most computationally intensive applications pre-install data before calculation. Compared with computational overhead, communication overhead is negligible. Therefore, this article does not consider the term frequency execution of the electric power terminology. In the term frequency model of electric power terminology in this article, network nodes adopt a space sharing mechanism, and each node in the term frequency terminology of electric power represents a computing sub-service [5]. Assume that there are \( m \) nodes \( \text{Node}_i, i = 0, 1, \ldots, m - 1, n \) in the data mining matching algorithm system. Each sub-service \( H_i, j = 0, 1, \ldots, n - 1 \). In order to realize the reasonable migration of workflow failure and recovery, each sub-service is assigned to a network node, and the following three random variables are used to describe the execution of the computing service, namely, service execution time \( \overline{T}_j^C \), service start time \( \overline{T}_j^S \) and the service end time \( \overline{T}_j^F \), and meet \( \overline{T}_j^F = \overline{T}_j^S + \overline{T}_j^C \) definition 2 (deadlock electric power term frequency service). If there is a non-empty service subset \( D \subseteq V \), \( D \) in the electric power term frequency service set \( V \), the order of each service in \( D \) Sub-services are included in this subset, namely: \( D \subseteq V = \{pre(t_1) \cup pre(t_2) \cup \ldots \cup pre(t_i), t_1, \ldots, t_i \in D\} \)

Then the electric power terminology frequency service is a deadlock power terminology term frequency service. We suppose that services \( h_i \) and \( h_j \) are assigned to the working node \( \text{Node}_i \) for execution and \( h_i \) is before \( h_j \). If there are no other services between \( h_i \) and \( h_j \), then the service is defined. The difference between the service start time \( \overline{T}_j^S \) of \( h_j \) and the service end time \( \overline{T}_j^F \) of service \( h_i \),
$\Delta = T_j^S - T_i^F$ is the idle period, and the working node is also called idle resource. If the reorder node set of service $h_i$ is $PN(h_i), h_j \in PN(h_i)$, when $h_j$ satisfies $T_j^F = \max_{h_k \in PN(h_j)} T_k^F$, define $h_j$ the key node that serves the node $h_i$. The key path is composed of a set of key nodes, and the starting point of the path is the entry node of the power terminology service, and the end of the path is the power terminology exit node.

The meaning of the key node can be understood as: in the pre-order node of service $h_i$, the node with the latest service end time is recorded as the key node. A series of key nodes constitute the critical path.

$$T_j^C = \omega + Y_1 + Y_2 + \ldots + Y_S + Z_1 + Z_2 + \ldots + Z_S$$

Among them, $Y_i (1 \leq i \leq S)$ is the network downtime, $Z_i (1 \leq i \leq S)$ is the network recovery time, then the average expectation and variance of the sub-service calculation time are respectively

$$E(T_j^C) = \frac{1}{1-\lambda_f \mu_f} + \lambda_f \mu_f \omega$$

$$V(T_j^C) = \left( \frac{\mu_f^2 + \sigma_f^2}{(1-\lambda \mu)^2} + \mu_c^2 + \sigma_c^2 + 2 \frac{\lambda_f \mu_c}{1-\lambda_f \mu_f} \right) \lambda_f \omega$$

The cumulative distribution function of the service execution time can be defined, and the calculation formula is as follows:

$$P(T_j^C \leq t) = P(T_j^C \leq t | S = 0) + P(T_j^C \leq t | S > 0)P(S > 0)$$

Given a power terminology term frequency tuple group $G = (V, E)$, where $v \in V$ and $(u, v) \in E$, then the power terminology term frequency is n-level power terminology term frequency if and only if:

1. $V = L_1 \cup L_2 \cup \ldots L_n \cup (L_i \cap L_j) = \emptyset, i \neq j$.

2. For each $(u, v) \in E$, where $u \in L_i$ and $v \in L_j$, must have $i > j$. Among them, L is a subset of nodes in the electric power terminology frequency graph, V is a node, E is a set of edges, u and v are examples of nodes if the start time $S_j$ of a sub-service is the maximum end time of its neighbouring predecessor nodes, the cumulative distribution function of the start time $T_j^S$ of the sub-service in the electric power terminology frequency graph can be expressed as $P(T_j^S \leq t) = \prod_{S_k \geq \nu(S_j)} P(T_k^F \leq t)$ where, $\nu(S_j)$ means that the sub-service is in the electric power The set of reorder nodes in the term frequency of the technical term [6]. Similarly, the cumulative distribution function of $T_j^C$ can be calculated using $V(T_j^C)$. Since $T_j^S$ and $T_j^C$ are independent variables, the probability density function $f_{T_j^F}(y)$ of the service end time $T_j^F$ can be calculated by the following formula:

$$f_{T_j^F}(y) = \int_0^y f_{T_j^S}(y-t)f_{T_j^C}(y)dt$$
Among them, $f_{STj}$ and $f_{CTj}$ represent the probability density of $T^S_j$ and $T^C_j$ respectively. When calculating the execution time of the hierarchical power terminology service, the service start, execution and end time of the uppermost sub-service are first calculated as the next sub-service time calculation the entire calculation process reaches the bottom of the power terminology frequency, and the end time of the sub-service is the execution time of the entire power terminology frequency service.

4. Design and implementation of management tool for electric power terminology database

4.1. System Overview
The electric power terminology database management tool is one of the components of the project team’s research topic “data integration system based on data service matching”, and the connection with other subsystems in the integrated system is shown in Figure 3.

![Figure 3. Framework diagram of a data integration system based on data service matching](image)

The other parts of the system are developed by the project team members. The focus of this article is to develop a convenient and intuitive electric power terminology database construction and management tool to realize the intuitive online editing and processing of the electric power terminology database, automatic reasoning and checking, and flexible and convenient Document management enables domain data integration and data sharing to rise to the semantic level.

4.2. System flow
1) We use the correlation calculation method to find the basic vocabulary in the power marketing field from the power marketing field corpus, and then randomly select a part of the corpus from the power marketing field corpus as the corpus to be trained, and then use each power marketing field in the corpus to be trained the basic vocabulary is the centre, and it is combined with other nearby vocabularies to form vocabulary strings of different lengths. The calculation of the mutual information between
adjacent vocabularies in the vocabulary string and the language characteristics of the vocabulary string will not only have higher relevance to the power marketing field, but the vocabulary that meets the characteristics of Chinese language is used as a professional vocabulary in the field of electric power marketing, and is marked in the training corpus to generate training corpus. Finally, the training corpus is trained based on the conditional random field method to obtain a professional vocabulary extraction model, and use the professional vocabulary extraction model Realize the extraction of power marketing related professional vocabulary from the power marketing field corpus, manually convert it into the corresponding English, and add it to the power marketing professional vocabulary. Figure 4 shows the system flow chart.

2) Collect power system related vocabulary and corresponding translations, and add them to the power marketing professional vocabulary database, which mainly includes power system vocabulary, general dictionary and power marketing vocabulary extracted from step 1).

3) Establish a two-level index structure for the professional vocabulary database of electric power marketing. Among them, the first-level index contains keywords and the next Chinese character index pointer, and the second-level index contains keywords and other string group pointers.

4) Retrieve the professional vocabulary database of electric power marketing, the order of retrieval is electric power system vocabulary, general dictionary, and vocabulary extracted in electric power marketing field. After the retrieval, the retrieval result is returned to the user.

5) Carry out the reprocessing of the power marketing corpus and the balance corpus, and use the Chinese lexical analysis system ICTCLAS of the Chinese Academy of Sciences to segment the power marketing corpus and the balance corpus. This is the basis for determining the basic vocabulary in the power marketing field. The corpus in the field of power marketing refers to the literature in the field of power marketing. The balance forecast refers to documents covering power, social security, and professional qualifications.

Figure 4. The corpus extraction flow chart of the electric power terminology database
6) Extract the corpus to be trained, and perform automatic labelling of the training corpus. First, based on the word segmentation in step 5), the domain correlation calculation method is used to find words with a higher probability in the field of power marketing, and use them as power basic vocabulary in the marketing field, and then randomly select 20% of the corpus from the power marketing field as the corpus to be trained, and then find all adjacent vocabulary strings that contain the basic vocabulary in the power marketing field and meet the requirements of mutual information and language characteristics, and mark them. It is the professional vocabulary of electric power marketing, and finally the marked training corpus is obtained. Said meeting the requirements refers to the introduction of stop words and part-of-speech combinations as punishment factors for professional vocabulary screening on the basis of mutual information.

7) The paper uses the conditional random field method in machine learning to analyse the labelled training corpus, including the internal composition of each professional vocabulary, the relationship between the vocabulary and the context, and the appropriate feature template is selected for deduction training to obtain a professional vocabulary extraction model. And extract the professional vocabulary related to power marketing or the professional vocabulary in the new corpus from the power marketing domain corpus according to the extraction model.

5. System test analysis
The Chinese-English bilingual computer professional term extraction system described in this article is implemented in Java language. Corresponding experiments and results analysis are carried out on the small English-Chinese parallel corpus "Test Corpus" that I have compiled and processed. "Test Corpus" contains 102 sets of aligned sentences, which is a small-scale English-Chinese bilingual electric power term corpus with sentence-level alignment. We reprocessed the corpus. For example, we performed morphological restoration in English and extracted candidate translation units using the N-gram method. A total of 1483 English candidate translation units and 468 Chinese candidate translation units were obtained. A total of 2926 candidate translation equivalent pairs were obtained. In order to improve the extraction speed, we first delete those candidate translation equivalent pairs that have a relatively small co-occurrence frequency. Here, because of the small size of our electric power term corpus, we only delete candidate translation equivalent pairs with a co-occurrence probability less than 2, and delete low frequencies [7]. After candidate translation equivalence pairs, there are 135 candidate translation equivalence pairs with a co-occurrence frequency greater than 2, and then the Dice coefficient, Phi square coefficient and Log Likelihood Ratio are used to calculate the translation probability. Finally, iterative method is used to extract translation equivalent pairs according to the order of translation probability from largest to smallest. According to Log Likelihood Ratio, a total of 55 English-Chinese translation equivalent pairs were extracted, of which 48 were completely correct and 5 were partially correct. We extracted a total of 56 English-Chinese translation equivalent pairs when we extracted the dice coefficients, of which 45 were completely correct and 6 were partially correct. According to Phi square coefficient extraction, the total number of English-Chinese translation equivalent pairs extracted is 57, of which 49 are completely correct and 4 are partially correct. See Table 1 for details.

| Translation probability calculation method | Dice coefficient | Phi square coefficient | LLR |
|------------------------------------------|------------------|------------------------|-----|
| Extract the number of correct translation equivalent pairs | 45               | 49                     | 48  |
| Extract the number of partially correct translation equivalent pairs | 6                | 4                      | 5   |
| The total number of translation equivalent pairs extracted by the system | 56               | 57                     | 55  |
| Accuracy of extraction                   | 0.8571           | 0.8947                 | 0.9182 |

Table 1. Dice coefficient, Phi square coefficient and translation equivalent evaluation results obtained by LLR
From the analysis of the data, it can be seen that the correct rate of extraction according to the Log Likelihood Ratio is significantly better than that of extraction according to the Dice coefficient and the Phi square coefficient. Moreover, the correct rate of extraction according to the method proposed in this article is significantly higher than that of traditional methods. This algorithm uses the N-gram model to solve the impact of multi-word correspondence and word segmentation errors on extraction, and it uses the iterative strategy of greedy assumption to effectively solve indirect related problems are eliminated, and a large number of wrong translation equivalents are eliminated. In addition, the method given in this article has found a lot of multi-word translation equivalence pairs under the three statistical measures. Most of these translation equivalence pairs are terms in the field. These terms or new words are usually mistaken in Chinese word segmentation. Is divided into multiple vocabulary units. Compared with traditional extraction methods, the N-gram model effectively recalls the translation equivalent pairs corresponding to these multiple words.

From the results in Figure 5 above, it can be seen that after calculating the overall feature values of the corpus, the Chinese and English corpora obtained account for 66.85% and 57.2% of the total number of the initial corpus respectively, but the comparison between the candidate comparable corpus and the sample corpus is similar [8]. After calculating the degree, Chinese corpus accounted for 51.57% of the total initial corpus, while the percentage of English corpus in the total initial corpus dropped rapidly to 36.09%. From the test results of the Chinese corpus, it can be seen that the experimental method has a certain degree of feasibility, with a comparable degree of over 50%, but the relevance of the obtained English corpus is not optimistic. The method based on seed words has certain limitations: the quality of seed words directly affects the quality of the generated comparable corpus. On the one hand, the current indexing keywords in some journal articles are not standardized, and there are cases of missing or wrong labelling; on the other hand, the document coverage of the seed words themselves is limited, and the number of seed words can be further expanded through the method of co-occurrence statistics.

6. Conclusion
Based on the professional vocabulary extraction method based on the combination of language characteristics and conditional random fields, we first find out the basic vocabulary in the power marketing field, and then randomly extract a part of the corpus from the power marketing field corpus as the corpus to be trained, and then based on the basic marketing vocabulary combined with mutual information. The language characteristic formula, extract the professional vocabulary in the power marketing field from the training corpus, mark these vocabularies in the training corpus, generate the training corpus, and finally train the training corpus based on the conditional random field method to
obtain the professional vocabulary extraction. The model based on this model can extract the professional vocabulary related to power marketing in the power marketing corpus or the professional vocabulary in other new corpora, which greatly improves the quality and efficiency of the extraction of professional vocabulary.

References
[1] Lahlou, H. The conceptualization of science terminology: A Cognitive linguistic analysis of the categories ELECTRICITY and LIGHT in Arabic. International Journal of Humanities and Social Science Research, 4 (2) (2018) 75 - 80.
[2] Wang, J., Wang, X., Ma, C., & Kou, L. A survey on the development status and application prospects of knowledge graph in smart grids. IET Generation, Transmission & Distribution, 15 (3), 383 - 407.
[3] Sasu, L. (2018). Challenges in Romanian/English Translation of Technical Terminology. Research and Science Today, 16 (2) (2021) 166 - 176.
[4] Balagiu, A., & Patesan, M. Homonymy within English Marine Engineering Terminology. Scientific Bulletin" Mircea cel Batran" Naval Academy, 21 (1) (2018) 1 - 8.
[5] Duffey, R. B. Power restoration prediction following extreme events and disasters. International Journal of Disaster Risk Science, 10 (1) (2019) 134 - 148.
[6] Smutný, M. Terminology as a Specific Carrier of Information. Prague Journal of English Studies, 7 (1) (2018) 143 - 160.
[7] Rezoug, F., & Vincent, B. Exploring Lexical Bundles in the Algerian Corpus of Engineering. Arab Journal of Applied Linguistics, 3 (1) (2018) 47 - 77.
[8] Lewis, K. False Friends in Terminology: Croatian Lost in Translation. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 46 (1) (2020) 141 - 164.