ELETerm: A Chinese Electric Power Term Dataset

Yi Yang, Liangliang Song, Shuyi Zhuang, Shi Chen, and Juan Li

State Grid Jiangsu Electric Power Company Research Institute, China

Motivation

To extract domain-specific terms, lots of existing studies have explored different methods. Some of them leverage the experience of experts and semantic rule based methods to design hand-crafted features and extract terms from domain-specific corpus, which requires sufficient labeled data, and large annotation costs. With the great success of deep learning models, some studies attempt to propose a term extraction model using the combination of Bi- directional Long Short-Term Memory (BiLSTM) and Conditional Random Field (CRF).

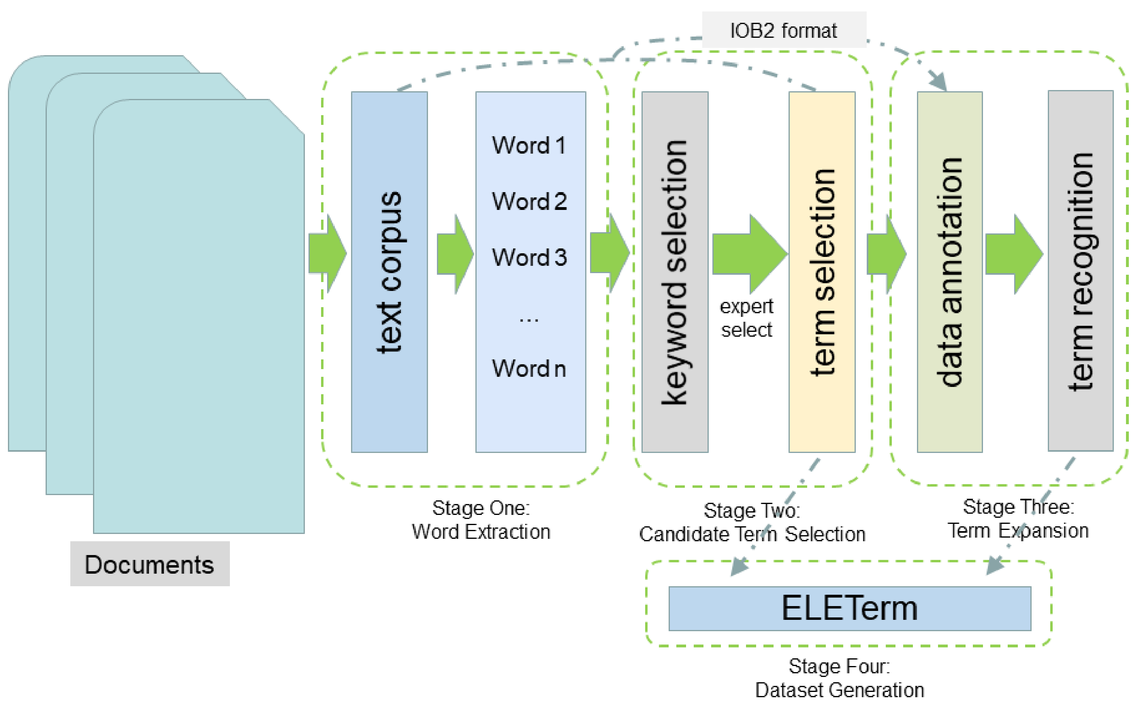
Although the previous studies have achieved good results, it still has the following problems, especially in the electric power field.

Problem1: the existing studies are usually used for method testing, and seldom generating large-scale domain-specific datasets;

Problem2: in the evaluation phase, it is hard for technologists to judge the correctness of the extracted terms since they are not familiar with domain-specific knowledge.

As a result, a high-quality dataset of electric power terms is still blank, which restricts the construction of professional knowledge graphs and their applications. To address this issue, we introduce ELETerm, a new Chinese term dataset, in which terms are extracted from business documents and professional documents.

Framework



Stage One: Word Extraction

We first extract words from text corpus. We employ a classic statistical feature based word extraction method to generate words from text corpus due to its good generalization ability in Chinese word extraction. In this stage, we can extract 32,758 candidate words.

Stage Two: Candidate Term Selection

1) **Keyword selection:** we adopt a voting method to select keywords from the extracted candidate words. We first use three methods, including RAKE, TextRank, YAKE, to extract keywords from the candidate words. We obtain approximately 6,100 keywords after this step.
2) **Term selection:** we invite electric power experts from State Grid Corporation of China to filter out the keywords (e.g., “人 工智能控制 (artificial intelligence control)”) which are irrelevant to the electric power field. Here, we obtain 4,328 accurate terms.

Stage Three: Term Expansion

1) **Data annotation:** we use the extracted terms to annotate the text corpus in the IOB2 format. We first segment the corpus into sentences, and remove the stop words. Then, we take the extracted terms as seeds to annotate the processed sentences. Specially, for each character in a given sentence, if it is the begin of a term, then it will be annotated ”B”; if it is the intermediate of a term, then it will be annotated ”I”;else it will be be annotated ”O”.

2) term recognition: we first use the annotated sentences to train a BiLSTM-CRF model which uses the Conditional Random Field (CRF) to output the optimal label sequences in the IOB2 format. Then, we send the unlabeled sentences as the input to the model, it will output the labels of sentences. We will get the term by selecting the characters which are corresponded to label ”B” and ”I”. After manual verification by experts, we obtain 5,484 terms to expand the dataset.

Stage Four: Dataset Generation

We use the terms obtained from stage two and stage three to build ELETerm, and this term dataset contains 10,043 terms in total.

1	光纤通道保护	248	自动电压	501	变引线
2	光纤通道	249	励磁系统	502	双极保护
3	继电保护	250	减磁装置	503	直流
4	电压等级	251	时钟信号	504	直流极保护
5	纵联	252	谐波分析	505	直流换流
6	保护通道	253	故障点	506	极引线
7	光纤通道传输	254	直流电源回路	507	直流双日极保护
8	电力调度	255	双跳闸绕组	508	滤波器保护
9	通信中心	256	短路暂态过程	509	锦屏站换流
10	电力	257	短路条件	510	直流换
11	电力系统	258	后备原则	511	交流滤波器
12	分相电流	259	电磁式	512	交流滤波母线保护
13	差动保护	260	铁磁谐振	513	控制功能
14	分相	261	交流电压回路	514	直流线路
15	电流	262	母差保护	515	接收装置
16	双重化	263	中性线	516	断路器跳闸接收装置
17	通道方式	264	纹波系数	517	断路器跳闸装置
18	开关	265	高电压	518	后断器
19	距离保护	266	接地网	519	安控装置
20	纵联保护	267	电位	520	直流功率
21	变电站	268	接地端子	521	交流断
22	保护装置	269	屏蔽层	522	定值管理

https://github.com/wuyike2000/ELETerm

Experiments

Table 1. Resources of ELETerm dataset.

Business documents	Equipment ledgers, Protection setting values Protection action information Device alarm information, Defect disposal Power grid events, etc.
Professional knowledge documents	Substation secondary drawings Installation instructions, Training courses Regulations, Technical standards Work instructions, Calibration reports Competition question bases

Table 2. The analysis result of terms in different categories includes the lexicology feature: the average word length (Ave-Len), the statistical feature: the average word frequency in text corpus (Ave-Wfreq), and the dictionary feature: the proportion of words in dictionary of common words (Common-Wrate).

	Noun	Verb
Numb	6,408	3,625
The lexicology feature **Ave-Len**	5.095	3.976
The statistical feature **Ave-Wfreq**	0.078	0.069
The dictionary feature **Common-Wrate**	0.114	0.132

Conclusion

In this paper, we introduce ELETerm, a Chinese domain-specific term dataset in the field of electric power. Based on high-quality resources, we carefully design a extraction approach to obtain a high-quality dataset. The statistics and analysis show the great prospect for ELETerm to further build electric power knowledge construction. The introduction of this dataset can help fill in the blank of the Chinese electric term dataset. As for the future work, we plan to continually expand this dataset and build large-scale electric power knowledge graph.