Global Energy Interconnection
Development and Cooperation Organization

Volume 3 Number 2 April 2020 (186-192)
DOI: 10.14171/j.2096-5117.gei.2020.02.010

Global Energy Interconnection

Contents lists available at ScienceDirect
https: //www.sciencedirect.com/journal/global-energy-interconnection

Full-length article

Power entity recognition based on bidirectional long short-term memory and conditional random fields

Zhixiang Ji[1], Xiaohui Wang[1], Changyu Cai[1], Hongjian Sun[2]

1. China Electric Power Research Institute Co. Ltd., Beijing, 100192, P.R. China

2. University of Durham, The Palatine Centre, Stockton Road, Durham, DH1 3LE, UK

Scan for more details

Abstract: With the application of artificial intelligence technology in the power industry, the knowledge graph is expected to play a key role in power grid dispatch processes, intelligent maintenance, and customer service response provision. Knowledge graphs are usually constructed based on entity recognition. Specifically, based on the mining of entity attributes and relationships, domain knowledge graphs can be constructed through knowledge fusion. In this work, the entities and characteristics of power entity recognition are analyzed, the mechanism of entity recognition is clarified, and entity recognition techniques are analyzed in the context of the power domain. Power entity recognition based on the conditional random fields (CRF) and bidirectional long short-term memory (BLSTM) models is investigated, and the two methods are comparatively analyzed. The results indicated that the CRF model, with an accuracy of 83%, can better identify the power entities compared to the BLSTM. The CRF approach can thus be applied to the entity extraction for knowledge graph construction in the power field.

Keywords: Knowledge graph, Entity recognition, Conditional Random Fields (CRF), Bidirectional Long Short-Term Memory (BLSTM).

1 Introduction

In recent years, with the development of smart grids, traditional grid facilities have been continuously upgraded. Furthermore, various information systems have been employed, and the use of smart grids has led to the generation and accumulation of large amounts of heterogeneous data from multiple sources. With the construction of the ubiquitous electric power Internet of Things, artificial intelligence based on the electricity big data is expected to play a key role in supporting the construction of professional applications such as marketing, operation inspection, material monitoring, dispatching, and security supervision. In addition, the in-depth development of artificial intelligence applications is expected to promote innovations in novel businesses, and new application models continue to emerge. However, at present, a domain knowledge graph for the power industry has not been established, and a fully intelligent electric power knowledge support system has not been realized.

Received: 22 December 2019/ Accepted: 23 February 2020/ Published: 25 April 2020

✉ Zhixiang Ji
jizhixiang@epri.sgcc.com.cn

Changyu Cai
caichangyu@epri.sgcc.com.cn

Xiaohui Wang
wangxiaohui@epri.sgcc.com.cn

Hongjian Sun
hongjian.sun@durham.ac.uk

The knowledge graph describes the concepts, entities, and their relationships in the objective world in a structured form. Specifically, this graph is a structured semantic knowledge base used to describe the concepts and their relationships in the physical world in a symbolic form. The basic constituent units are "entity-relation-entity" triples, and the entities and their related attribute-value pairs. The entities are interconnected through relationships to form a networked knowledge structure. In the field of dispatching, the knowledge graph can be used to solidify the dispatching procedures and dispatcher experience and knowledge and support applications such as power grid operation monitoring, exception handling, and mode adjustment. In the field of operation inspection, the knowledge graph can be used to store equipment, faults, and disposal methods, and it can support intelligent operation and maintenance of equipment.

To construct a domain knowledge graph, first, the power entities are extracted. Entity recognition is a basic task in natural language processing and has a wide range of applications. Domain-specific named entity recognition, the goal of which is to identify domain-specific entities and their categories, plays a significant role in domain document classification, retrieval, and content analysis. This task is the basis of deep and complex information extraction tasks and the cornerstone of the knowledge calculation process that transforms data into machine-readable knowledge.

In recent years, with the increasing demand for knowledge graphs, many industries have attempted to realize entity recognition for specific fields. Biomedical entity recognition for the medical field has been realized, and a combination of support vector machines, conditional random fields (CRF), and rules has been used to identify medical entities and institutions [1-3]. Furthermore, a combination of CRF and rules has been adopted in the military and geographic domains to realize named entity recognition [4-6].

With the development of neural networks, especially those for deep learning, neural-network-based methods have been applied for Chinese word segmentation [7, 8], sequence labeling, and entity recognition [9-12] and achieved satisfactory results.

In particular, the texts in different fields have different characteristics. Therefore, to realize power entity recognition, it is necessary to refer to the entity recognition technologies in other fields and optimize the methods. The power domain involves several entities such as power equipment and concepts, and multiple abbreviations or acronyms, as well as multiple names of the same entity may be present. Nevertheless, the power field does not yet have a mature power text corpus, and there is a lack of labeled data for machine learning. Therefore, entity recognition in the power domain is a challenging task. In this work, CRF and bidirectional long short-term memory (BLSTM) models are used for entity recognition, and a comparative test is performed.

2 Power entity recognition

2.1 Named entity recognition

Named entity recognition is a basic task of natural language processing. This task is to label named entities in unstructured text [13]. Named entities generally refer to entities with specific meanings or strong references in the text, generally including the names of people, places, organizations, dates, time, and proper nouns. These entities are extracted from unstructured input text and named entity recognition is performed to identify more categories of the entities based on the business needs. Furthermore, named entity recognition is the basis of relationship extraction, and named entities are the research subjects of entity recognition. Generally, named entities can be classified into three categories (entity, time, and number) or seven categories (person, place, institution, time, date, currency, and percentage).

2.2 Power entity recognition

In actual research pertaining to specific fields, the exact meaning of the named entities needs to be determined according to the specific applications. When constructing a power knowledge graph, the power entities include traditional named entities as well as entities such as the power equipment, accessories, substations, and transmission lines. The typical entities are listed in Table 1.

2.3 Challenges to power entity recognition

At present, most entity recognition algorithms use supervised learning, which requires a large amount of labeled data, and text labeling data in the power field is presently lacking. In particular, various types of power entities and a high degree of professionalization is present in the power domain, and a large number of abbreviations, aliases, and nesting is expected to be present. In such a case, the deformation form is difficult to identify, which constitutes a technical challenge to the recognition algorithm. With an increase in the entities in the power sector and their nicknames with the advancement of business and technology, the new text is expected to include more power entities. Furthermore, several nonexistent entity words need professional recognition. This requires the algorithm model to have a strong recognition ability.

Table 1 Sample power entity

Entity type	Entity description	Entity instance
Name	Dispatchers, inspectors, etc.	Li Fei
Organization	Power company or organization	Shandong Electric Power Group Corporation, Inspection Class one
Geography	Place name	Beijing
Device	Device name or abbreviation	Transformer, circuit breaker, protection device, fault location device
Substation	Substation name	220 kV Fuzhou Substation, 500 kV Pailou Substation
Line	Line name	Ansong double circuit line, 220 kV brand Yang I line

3 Power entity recognition technology

In the early work pertaining to named entity recognition, methods based on artificial construction rules were usually used. With the increase in data size and computational power, many supervised machine learning methods are available for application. Such methods are based on a large number of labeled field texts. The positive and negative examples are learned to construct a machine learning model to realize the labeling and recognition of the entities.

3.1 Rule-based approach

Rule-based methods use corpus analysis and incorporate the linguistic knowledge to manually construct rules to identify the entities. Furthermore, these methods match the text with manually constructed rules to identify the entities during the recognition task [14,15]. For power texts, "disconnect" can be used as the following term for equipment such as "circuit breakers" and "knife brakes", and words such as "transformation" and "station" can be used as the end phrase for a transformer, substation, etc. When constructing rules, linguistic knowledge such as part-of-speech and syntax can be used. Because the linguistic knowledge is complex, several conflicts among the rules must be handled. Consequently, although rule-based methods are effective in some cases, building the rules is time-consuming and labor-intensive, and the portability of the approach is poor.

3.2 Machine-learning-based approach

The existing named entity recognition methods generally convert the named entity recognition problem into a sequence labeling problem and then solve it via sequence labeling, in which the problem is converted into a structured classification problem.

In a labeling problem, given a sequence of input text, the model can output the labeling result of the original sentence. The model needs to be learned through a large amount of labeled corpus training data, and the model can predict the labeling results based on the input observation sequence. The commonly used entity recognition models include hidden Markov models (HMM) and CRF. [16-18]. Among the existing schemes, the most commonly used scheme is the feature template and CRF scheme, which constructs binary features via the manual definition function and mines the internal characteristics of the named entities and the context characteristics. The location of the feature is a window, which is also the context location. The different feature templates can be combined to form a new feature template.

4 Power entity recognition based on BLSTM and CRF

4.1 CRF model

The CRF model is a discriminative model that combines the characteristics of the maximum entropy model and HMM model. The approach can yield satisfactory results when tagging natural language processing sequences such as part-of-speech tagging and entity recognition [19,20].

The CRF model is a statistical model based on a probability map, which can use the context information to incorporate a variety of features and implement global normalization. Furthermore, the approach can obtain a global optimal solution and solve the problem of label deviation.

Generally, $X=\{x_1, x_2, \ldots, x_n\}$ is used to represent an input sequence, where x_i represents a vector of the i-th input character, and $y=\{y_1, y_2, \ldots, y_n\}$ represents a possible label sequence of input x. $Y(x)$ represents a possible tag sequence for X. The sequence CRF probability model defines a series of conditional probabilities $P(y|z, W, b)$. The probability of all the possible sequences y given x is calculated using (1):

$$p(y|z;W,b)=\frac{\prod_{i=1}^{n}\varphi_i(y_{i-1},y_i,z)}{\sum_{y'\in Y_{(z)}}\prod_{i=1}^{n}\varphi_i(y'_{i-1'}\ y'_{i'}z)} \tag{1}$$

where $\varphi_i(y',y,z)=\exp(W^{T}_{y',y}{}^{zi}+b_{y',y})$ is a potential function; $W^{T}_{y',y}{}^{zi}$ and $b_{y',y}$ are the weight vectors for the label pair (y', y) and the offset, respectively.

The maximum likelihood estimation technique is used to train the CRF model. For the training set {(zf, Yf)}, the likelihood estimation algorithm can be expressed as

$$L(W,b)=\sum log\, p(y|z;W,b) \tag{2}$$

In the maximum likelihood training process, parameters to maximize L (W, b) are selected. The decoding process involves identifying the label sequence y^* with the highest conditional probability, as shown in (3).

$$y^* = argmax_{\,y\in Y(Z)}\, p(y|z;W,b) \tag{3}$$

In the sequence controlling instruction understanding model, only the interaction between two consecutive labels is considered, and the Viterbi algorithm can be used to achieve excellent results in the training and decoding process.

4.2 BLSTM model

The long short-term memory (LSTM) represents a special recurrent neural network (RNN), which solves the problem of the vanishing gradient of the RNN by introducing a memory unit and threshold mechanism [21]. The structure of an LSTM unit is shown in Fig. 1, where x represents the input of the network at different times, y is the output of the network, h denotes the hidden layer, u is the weight from the input layer to the hidden layer, w is the weight of the previous node hidden layer to the the current node hidden layer, and v is the weight from the hidden layer to the output layer.

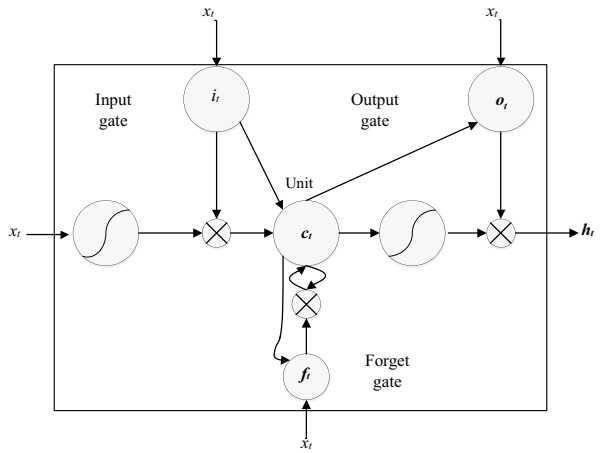

Fig. 1 LSTM unit structure

In the specific implementation of the LSTM, the LSTM unit is updated at time t using the following expressions:

$$i_t=\sigma(w_i h_{t-1}+U_i x_t+b_t) \tag{4}$$

$$f_t=\sigma(w_i h_{t-1}+U_f x_t+b_f) \tag{5}$$

$$\widetilde{c}_t=tanh(W_c h_{t-1}+U_c x_t+b_c) \tag{6}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\widetilde{c}_t \tag{7}$$

$$o_t=\sigma(w_o h_{t-1}+U_o x_t+b_o) \tag{8}$$

$$h_t=o_{t-1}\odot tanh(c_t) \tag{9}$$

Here, σ is the sigmoid function, and Θ is the corresponding product of the elements. x_t is the input vector at time t. h_t is the hidden state vector, which is also known as the output vector and stores all the information at time t and the previous time. U_i, U_f, U_c, U_o are the weights of the input vector xt for the input gate, forgotten gate, unit gate, and output gate, respectively. W_i, W_f, W_c, W_o are the weights of different gates to the hidden state vector h_t. b_t, b_f, b_c, b_o are the offset vectors. By employing the three gate structures, the LSTM enables the recurrent network to retain the useful information for the tasks in the memory unit during the training process, thereby avoiding the problem of the RNN disappearing when acquiring long-range information.

The structure of the BLSTM is as shown in Fig. 2. When processing sequence data, the BLSTM introduces an additional backward calculation process, which is different from the ordinary LSTM case. This process can use the following information of the sequence. Finally, the forward and reverse calculations are performed. The values are output to the output layer simultaneously; consequently, all the information of a sequence is obtained through two directions, which can be applied to multiple types of natural language processing tasks [22-25].

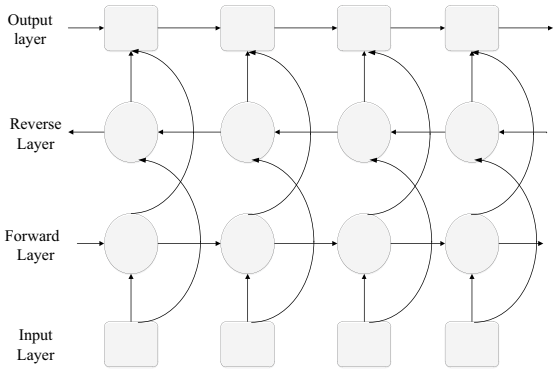

Fig. 2 BLSTM model

5 Power entity recognition experiment

5.1 Corpus annotation

Because only few corpora of power texts are available, and none of the corpora are marked, in the experiment, the corpora are manually constructed. Specifically, power science and technology papers are used as the corpora. The entity is determined via manual verification, and the corpus size is 19469 sentences.

The corpus annotation uses the BIEO annotation set, where B-electric, I-electric, and E-electric respectively represent the first, internal, and final characters of the power

entity, and O represents that the character is not a part of the named entity.

5.2 Training model

The labeled data is used as the corpus required for the entity recognition supervision training. In the experiment, the corpus data is divided into five parts, and the number of training data points is 15648, 15800, 15695, 15918, and 15576. Each remaining corpus is used as test data, with 3819, 3667, 3772, 3549, and 3891 data points. Experimental verification is performed, and five models are generated.

To accelerate the model learning, the neural network model adopts the Adam optimizer. The parameters of the model are listed in Table 2.

Table 2 Parameter setting for BLSTM training

Parameter	Value
batch_size	64
epoch	10
hidden_dim	300
optimizer	Adam
learning rate	0.001
gradient clipping	5.0
dropout keep_prob	0.5
update embedding	true
pretrain embedding	random
embedding_dim	300
shuffle	true

5.3 Evaluation index

The evaluation indicators used in power entity recognition mainly include the accuracy rate and recall rate, which can be defined as in (10) and (11), respectively.

$$\text{Recall}=\frac{\text{Number of correct entities identified}}{\text{Number of entities in the sample}} \tag{10}$$

$$\text{Accuracy}=\frac{\text{Number of correct entities identified}}{\text{Number of correct identified entities}} \tag{11}$$

The values for both these indicators lie between 0 and 1. A value closer to 1 corresponds to a higher accuracy or recall. The accuracy rate and recall rate contradict each other in some cases. Therefore, it is necessary to comprehensively consider their weighted harmonic average value, that is, the F value, which can be defined as in (12).

$$F=\frac{2*\text{Accuracy}*\text{Recall}}{\text{Accuracy}+\text{Recall}} \tag{12}$$

5.4 Experimental results and analysis

In the experiment, the BLSTM model and CRF model were used for power entity recognition, and the corresponding results are presented in Tables 3 and 4, respectively.

Table 3 Experimental results for the BLSTM model

No	Accuracy (%)	Recall (%)	F value (%)
1	75.81	82.93	79.21
2	75.01	82.58	78.62
3	75.26	81.69	78.34
4	75.71	82.89	79.13
5	75.61	82.70	79.00

Table 4 Experimental results for the CRF model

No	Accuracy (%)	Recall (%)	F value (%)
1	83.07	81.89	82.48
2	82.96	81.61	82.28
3	83.49	81.98	82.73
4	82.63	81.50	82.06
5	84.00	82.34	83.34

The abovementioned initial results are weighted for the comparative analysis, as shown in Table 5.

Table 5 Comparison of model results

Model	Accuracy	Recall	F value
BLSMT	0.7548	0.82558	0.7886
CRF	0.8323	0.81864	0.82578

The experimental results indicate that the CRF model can better identify the power entities compared to the BLSTM model.

The results of some of the extracted entities are incorrect, mainly because the corpus of this training model involves sentences extracted from power-related papers. The sentences are relatively professional and lack nonprofessional literature data. However, some nonprofessional entities may affect the accuracy of the model.

Because the corpus used in the experiment involves approximately 20,000 sentences, the corpus size is relatively small, which limits the model learning ability. This problem can be overcome by expanding the corpus size used for training.

6 Conclusions

(1) Entity recognition is the basis for constructing knowledge graphs in the power field. The CRF model can satisfactorily extract the power entities and thus effectively support the construction of power knowledge graphs.

(2) The corpus is the basis for mining and constructing of electric power knowledge. To support the application of artificial intelligence in various fields of electric power, the electric power industry needs to integrate the relevant resources to build a unified electric power tagging corpus.

(3) As a form of knowledge representation, the knowledge graph is expected to play an increasingly important role in the fields of dispatching, operation inspection, and intelligent customer service, thereby effectively supporting various applications of the ubiquitous electric power Internet of Things.

Acknowledgements

This work was supported by Science and Technology Project of State Grid Corporation(Research and Application of Intelligent Energy Meter Quality Analysis and Evaluation Technology Based on Full Chain Data).

References

[1] Wang H, Zhao T (2006) Recognition of named biomedical entities based on SVM. Journal of Harbin Engineering University 27: 570-574

[2] Li W, Zhao D, Li B, et al (2015) Recognition of medical medical record entities based on CRF and rule. Journal of Computer Applications 32(4): 1082-1086

[3] Zhang J, Wang S, Qian C (2014) Identification of Chinese medical institution names Based on CRF and rules. Journal of Computer Applications and Software 3: 159-162

[4] Feng Y, Zhang H, Hao W (2015) Named entity recognition for military texts. Computer Science 42(7): 15-18

[5] Jiang W, Gu J, Cong L (2011) Research on military named entity recognition based on CRF and rules. Command Control and Simulation 4: 19-21

[6] He Y, Luo Ci, Hu B (2015) Geographic named entity recognition method based on the combination of CRF and rule. Journal of Computer Applications and Software 32(1): 179-185

[7] Chen X, Qiu X, Zhu C, et al (2015) Gated recursive neural network for Chinese word segmentation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, July 2015, vol 1, Beijing, China, pp. 1744-1753

[8] Chen X, Qiu X, Zhu C Liu P, Huang XJ (2015) Long short-term memory neural networks for Chinese word segmentation. 2015 Conference on Empirical Methods in Natural Language Processing, September 2015, Lisbon, Portugal, pp. 1197-1206

[9] Wu Y, Jiang M, Lei J, Xu H (2015) Named entity recognition in Chinese clinical text using deep neural network. Studies in health technology and informatics 216: 624-628

[10] Santos CN, Guimaraes V (2015) Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv: 1505.05008

[11] Chen D, and Manning CD (2014) A fast and accurate dependency parser using neural networks. Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing, October 2014, Doha, Qatar, pp. 740-750

[12] Chiu JPC, Nichols E (2016) Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4: 357-370

[13] Zhang X, Wang T, Chen H (2005) Research on named entity recognition. Computer Science 32(4): 44-48

[14] Zhou K (2010) Research on rule-based named entity recognition. Dissertation, Hefei University of Technology

[15] Cheng Z (2015) Research on Chinese named entity recognition method based on rules and conditional random fields. Dissertation, Huazhong Normal University

[16] Han P, Jiang J (2010) Application research of HMM in the field of natural language processing. Computer Technology and Development 20(2): 245-248

[17] Feng Y, Yu H, Sun Geng, et al (2016) Method of domain term recognition based on word vector and conditional random field. Journal of Computer Applications 36(11): 3146-3151

[18] Xie Z (2017) Research on Chinese named entity recognition algorithm. Dissertation, Zhejiang University

[19] Shi M, Li B, Chen X (2010) Research on integration of pre-Qin Chinese word segmentation based on CRF. Journal of Chinese Information Processing 24 (2): 39-46

[20] Zhang J, Wang S, Qian C (2014) Identification of Chinese medical institution names based on CRF and rules. Journal of Computer Applications and Software 3: 159-162

[21] Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8): 1735-1780

[22] Cheng C, Hong T, Xue S (2019) BLSTM_MLPCNN model for short text classification. Computer Science 6: 206-211

[23] Rekia K (2018) CCG hyper-annotation based on deep learning model. Dissertation, Harbin Institute of Technology

[24] An C, Huang J, Chang S, Huang Z (2016) Question similarity modeling with bidirectional long short-term memory neural network. 2016 IEEE First International Conference on Data Science in Cyberspace, June 2016, Changsha, Hunan, China, pp. 318-322

[25] Ray A, Rajeswar S, Chaudhury S (2015) Text recognition using deep BLSTM networks. 2015 eighth international conference on advances in pattern recognition, January 2015, Kolkata, India, pp. 1-6

Biographies

Zhixiang Ji received the master's degree at Harbin Institute of Technology, Nangang District, Harbin, 2011. He is working in China Electric Power Research Institute Co. Ltd., Haidian district, Beijing. His research interests include the application of artificial intelligence technology in power systems.

Xiaohui Wang received the Ph.D. degree at North China Electric Power University, Beijing, 2012. He is working in China Electric Power Research Institute Co. Ltd., Haidian district, Beijing. His research interests include power big data technology, artificial intelligence, active distributed networks, and energy internet.

Changyu Cai received the bachelor's degree and master's degree at Changchun University of Science and Technology, Changchun, Jilin, China, 2006 and 2010, respectively. He is working in China Electric Power Research Institute Co. Ltd., Haidian district, Beijing. His research interests include artificial intelligence and big data.

Hongjian Sun (IEEE, S'07-M'11-SM'15) received the Ph.D. degree in Electronic and Electrical Engineering from the University of Edinburgh, U.K., in 2011. He held post-doctoral positions with King's College London, U.K., and Princeton University, USA. Since 2013, he has been with the University of Durham, U.K., as a Reader in Smart Grid (with a Lecturer position in 2013-2017).

(Editor Zhou Zhou)