Automatic MEP Knowledge Acquisition Based on Documents and Natural Language Processing

Shuo Leng[1], Zhen-Zhong Hu[1,2*], Zheng Luo[3], Jian-Ping Zhang[1] and Jia-Rui Lin[1]
[1] Department of civil engineering, Tsinghua University, Beijing, China
[2] Graduate school at Shenzhen, Tsinghua University, Shenzhen, China
[3] Glodon Technology Inc., Beijing, China
* email: huzhenzhong@tsinghua.edu.cn

Abstract

Mechanical, Electrical and Plumbing (MEP) systems are critical assets in buildings. A series of systematic specifications have been developed and extensive experiences have been accumulated as human knowledge to guide the design and maintenance of MEP systems. However, most of the MEP-related knowledge is represented in the form of unstructured texts and heterogeneously dispersed in the design documents and Internet. It is therefore difficult for managing, querying and utilizing them. To address this challenge, the research study described in this paper constructed a knowledge graph by automatic collecting and storing of MEP knowledge from unstructured data. Specifically, the MEP documents were first acquired from the Internet, and multiple Natural Language Processing (NLP) techniques were then adopted to extract entity and discover relationship from the information documented in these documents. Finally, the knowledge graph was established and presented in a vivid form. The constructed knowledge graph is expected to contribute to the promotion of AI technology in the Architecture, Engineering and Construction (AEC) industry.

Keywords: MEP, Knowledge graph, NLP

1. Introduction

Modern Mechanical, Electrical and Plumbing (MEP) systems consist of multiple subsystems such as energy management, Heating, Ventilation and Air Conditioning (HVAC), water supply, lighting and emergency alarms, etc. (Hu, Tian, Li, & Zhang, 2018). With the development of the MEP engineering, knowledge has been accumulated in the community to improve system efficiency and reduce failures. Generally, knowledge is defined as a collection of descriptions, relationships, and processes in specific domains (Zhou, 2010). In the MEP domain, knowledge is described in industry specifications, technical manuals, research literature and encyclopedias. These documents specify the characteristics, attributes, and operation and inspection methods of MEP facilities, and describe how these facilities work together in a system. The knowledge guides building engineers to design and operate MEP systems in an efficient and reliable way.

The MEP-related knowledge is primarily written by design and operation professionals using

human language and presented in the form of text documents. Although these unstructured texts can be easily understood by humans, computers cannot acknowledge their meanings other than character codes. As a result, this data is only stored in computer systems for people to manually access. Meanwhile, as the knowledge is massively and heterogeneously dispersed on the Internet, information from different sources may be duplicated or conflicted, and errors may exist, which makes it difficult to manually collect and organize this knowledge in a systematic way. Junior MEP engineers would have difficulties in learning the knowledge. Moreover, the lack of high-quality, structured data has limited the Artificial Intelligence (AI) applications such as intelligent question-answering, and knowledge search and inference in the MEP field.

To address this challenge, the research study discussed in this paper constructed a knowledge graph by automatic collect and store MEP knowledge from the Internet. The knowledge graph is designed to organize entities and their relationships in the form of a semantic network, in which entities refer to concepts, terms, or anything else that can be called knowledge (Paulheim, 2017). With this structure, the entity can be accurately positioned, and things related to it can be quickly detected. The knowledge graph can be further used as databases for knowledge representation, search, inference and data mining, and has achieved successful applications in some industries (H. Wang, Miao, & Yang, 2018).

In order to construct the knowledge graph, text documents were first obtained from the internet, and multiple Natural Language Processing (NLP) techniques were applied. The structure of the paper is organized as follows: Section 2 investigates relevant literature on intelligent MEP and knowledge engineering. Section 3 introduces some necessary techniques in achieving automatic MEP knowledge acquisition. The next section gives a case study to present the constructed knowledge graph and evaluates its performance. Finally, the discussion and conclusion are given.

2. Related Research Studies

The application of computers and automated methods to improve the traditional construction and management process has been one of the researches focuses in the MEP field. Through a specially designed algorithm, the computer can detect the potential spatial conflicts of MEP components (L. Wang & Leite, 2016), or help designers detect drawing errors according to the design rules (Korman, Fischer, & Tatum, 2003). Building Information Modeling (BIM) technology can also be applied to the MEP field to improve the efficiency of facility management (Hu, et al., 2016). Though these studies provided automated solutions for those heavy repetitive tasks, the MEP knowledge they need, such as the rules of collision detection or the initial database for reasoning, still had to be manually collected and imported, which was time-consuming and limited further development of automation technology.

Establishing a unified and integrated knowledge base is a feasible way to solve knowledge problems. In the Architecture, Engineering and Construction (AEC) industry, the BIM-based semantic web is an effective method for knowledge management and representation (Pauwels, et al., 2017). The semantic web, which is a concept extended by the World Wide Web (WWW), is used to fuse and store information from different data sources by defining ontologies and their relationships (Berners-Lee, et al., 2001). Through the semantic web, BIM can integrate different aspects of information such as energy consumption (Jiang, et al., 2018) and be further applied to knowledge reasoning (Bouzidi, et al., 2012). However, in the MEP field, works related to semantic web have not been seen in the literature. Even in the AEC field, the construction of the semantic web mainly follows a top-down process, which means

that the semantic web needs to first determine the top-level concepts by domain experts, and then divide the instances into the corresponding concepts (Turk, 2006). A large amount of manual participation lead the semantic web limit in size, and the fixed top-level concepts also make it difficult for the database to be modified with the rapid update of knowledge in the network.

The term knowledge graph was originally proposed by Google in 2012 to name their semantic web products for search engines, and then extended to a concept referring to the web-based semantic web databases (Paulheim, 2017). In addition to the Google Knowledge Graph, some instances of general knowledge graph include DBpedia (Lehmann et al., 2015) and YAGO (Fabian, Gjergji, & Gerhard, 2007) were also presented in the common field. In the AEC industry, a geotechnical knowledge graph has been proposed and established (C. Wang, et al., 2018). Compared with the traditional semantic web, these knowledge graphs are equipped with information acquisition and processing algorithms and can be expanded automatically by extracting knowledge from the network. As a result, the knowledge graph is generally larger in scale and requires less human involvement.

Based on the knowledge graph, many successful applications of AI techniques have been achieved. For example, knowledge graphs can be used as databases for Q&A systems(Sawant, et al., 2019). Entities in the question statement can be extracted by decoding techniques, and the information of the entity in the graph can be provided to the user as answers. Knowledge graphs can also be used to support recommender systems (Wang, et al., 2018). The potential interests of the user can be extended as relationships among entities. When an entity is selected, other entities associated with it will be recommended to the user. A knowledge graph in the MEP field can promote such AI applications in the industry, providing convenience for MEP industry practitioners.

3. Primary Approaches

3.1 Data acquisition and preprocessing

The data used in the paper were derived from text documents (in the form of PDF, Word, and HTML) related to the MEP domain from the Internet. A crawler program was developed in this research to obtain documents from some typical MEP websites. Then these data were merged together to form the raw dataset for the study.

During the preprocessing process, data cleaning and format conversion were first performed. Then the text segmentation process, which divided a sequence of words into separate words, was executed. For English texts, words are always separated by spaces, and the segmentation is naturally completed. However, in Chinese, there is no obvious separation between words, and the process of text segmentation is thus indispensable (Wu & Tseng, 1993).

3.2 MEP Entity Extraction

The target of entity extraction is to automatically discover the MEP related entities in the text. This process utilizes the Named Entity Recognition (NER) technology. The entities here refer to MEP terms include facilities, operations and attributes, and they will participate in the construction of knowledge graph as network nodes. Several types of practical NER techniques have been proposed, and the model Bidirectional Long Short Term Memory (Bi-LSTM) network with a Conditional Random Field layer (Bi-LSTM-CRF) which has been proved to have a considerably good performance (Z. Huang, et al.,

2015) was applied to implement this process. The Bi-LSTM-CRF method converts the entity recognition process into a sequence labeling task and outputs the labeling results with the highest occurrence probability. The model structure of the Bi-LSTM-CRF is shown in Figure 1.

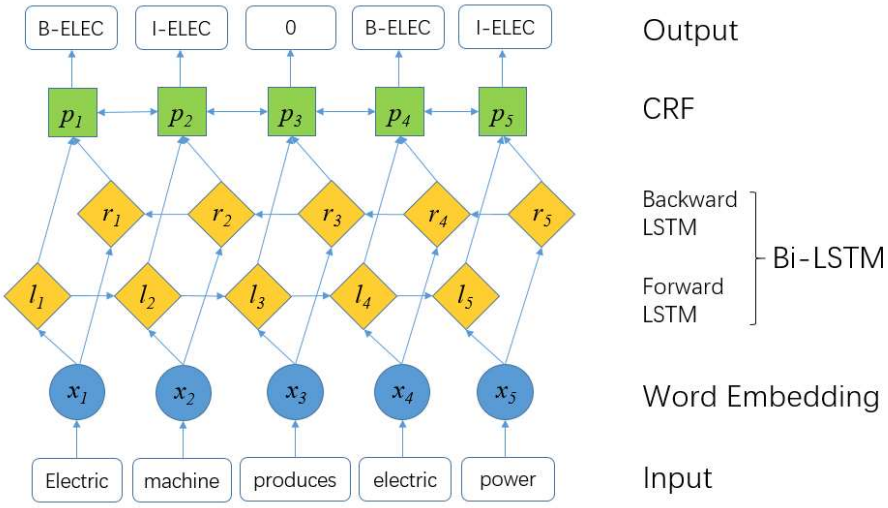


Figure 1: Structure of the Bi-LSTM-CRF model

The Bi-LSTM-CRF model consists of three parts: the word embedding layer, the Bi-LSTM layer, and the CRF layer. The model receives the word sequence in natural language as input variables and outputs a predictive label for each word after a series of calculations. Specifically, the word embedding layer converts each word into a vector, and the Bi-LSTM layer maps the vector to the probability of each label. Subsequently, the CRF layer adjusts the likelihood of the label based on probability theory, and the category with the highest probability is selected as the output label. Words with some specific labels are eventually determined to be named entities.

3.3 Entity Relationship Discovery

The entity relation discovery process is used to obtain the relationships between entities that act as direct edges in the knowledge graph, connecting entity nodes to form a network structure. These relationships show how the MEP entities connected to each other logically and work together. Multiple algorithms from rule-based methods to deep learning have been proposed for relationship discovery in this research. A Residual Convolutional Neural Network (ResCNN) model was adopted for its high accuracy and fast training speed (Y. Huang & Wang, 2017). The structure of ResCNN is illustrated in Figure 2.

Similar to the entity discovery process, each word needs to be converted to a vector at the beginning of the ResCNN model. Vectors of the entire sentence are then concatenated to a matrix and imported to the core model. The structure of the core ResCNN is basically the same as that of general CNN, including the convolution layer, pooling layer, full connection layer, and Softmax process. However, ResCNN adds a shortcut between every two or three convolutional layers so that the input value can directly participate in the calculation of the predicted result. Such a structure is called a residual block. In this way, instead of the predicted value, the learning target of the network focuses on the residual that is obtained by subtracting the input value from the predicted value. The final output of the model is the

probability of each relationship, and the category with the highest probability is selected as the relationship between the two entities.

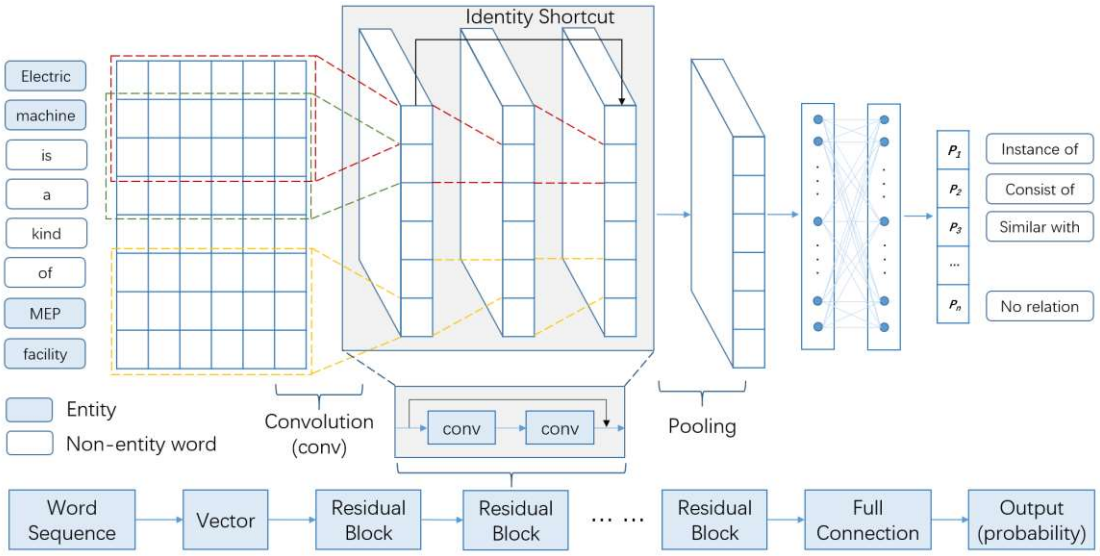


Figure 2: Structure of the ResCNN model

4. Case Study

In this paper, MEP text documents were collected from multiple websites on the Internet. The data consists of sub-datasets including industry specifications, research literature, encyclopedia, Q&A corpus and MEP component library. The raw text is 70M in size and contains more than 14 million words.

As a data preprocessing step, the word segmentation was performed to split the words from sentences. In the proposed approach, the text segmentation tool Jieba (Fxsjy, 2019) was applied to complete this process. The split words were then used as input variables for subsequent NLP models. Among them, 40% of the corpus was randomly selected for model training, and the rest was used for entity extraction after the model was fully trained.

Since machine learning methods convert entity recognition into tasks of sequence labeling, characters in the text need to be labeled before the algorithm is executed. In this research, the labeling method marked with BIO was adopted. The notation B represents the starting character of an entity, I represent the subsequent character of an entity, and O indicates a non-entity character. Three categories of entities were identified in the proposed approach including electric, HVAC, and water supply and drainage. The introduction and examples of the labeling method are shown in Table 1 and Figure 3 respectively. A dictionary mapping method which automatically labels texts by judging whether the word is in a pre-prepared MEP dictionary was adopted in this paper.

The marked text sequences were input to the Bi-LSTM-CRF model for training, and the fully trained model could then be used to predict the label for new texts as the NER process. 15 epochs were performed before the final prediction model was obtained. The predicted results of the model on the test dataset are illustrated in Table 2.

Table 1: Introduction to the sequence labeling method

Label	Explanation
B-E	Beginning character of an electric entity
B-H	Beginning character of an HVAC entity
B-W	Beginning character of a water supply and drainage entity
I-E	Subsequent character of an electric entity
I-H	Subsequent character of an HVAC entity
I-W	Subsequent character of a water supply and drainage entity
O	Characters that do not belong to entities

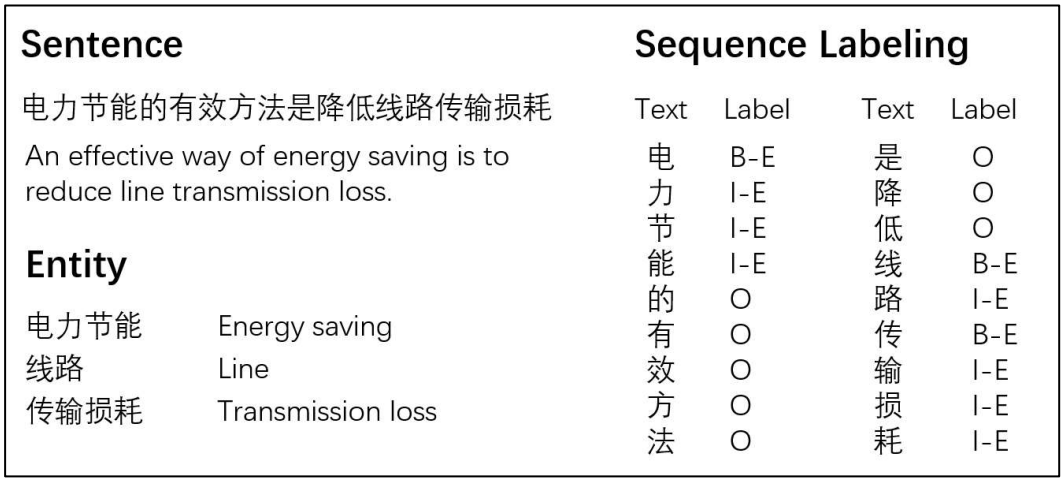


Figure 3: Instance of sequence labeling

Table 2: Training results of the Bi-LSTM-CRF model

Entity categories	Precision	Recall	F1 Score
Electric	97.01	97.97	97.49
HVAC	96.48	97.91	97.19
Water supply and drainage	96.15	97.53	96.84
Total	96.78	97.90	97.34

As can be concluded, the overall accuracy of the model is over 97%, which is a relatively high value compared with other state-of-the-art NER models. After the end of the NER program, a manual screening process was further executed to delete the entities found by mistake and improve the quality of the entity dataset. A total of 11332 MEP entities were eventually extracted.

After the NER process, the discovered entities were marked in the corpus and the relationship discovery was subsequently performed. As illustrated in Table 3, 4 types of relationships, including “have attribute”, “instance of”, “similar to” and “contain”, were defined to represent interactions between MEP entities. The relationship “instance of” could be judged by syntactic analysis, and the other categories were evaluated by the ResCNN model. Sentences in the corpus that contain two or more entities were selected as datasets for relationship discovery. Among them, relationships of entity pairs in 900 sentences were manually labeled to train the model, and the remaining sentences were used

for relationship discovery.

Table 3: Introduction to the relation category

Relationship categories	Explanation	Example
Have attribute	"Entity A has attribute Entity B" means that B is an attribute of A	A: air conditioning B: power
Instance of	"Entity A is an instance of Entity B" means that B is a general concept and A is a special case of B	A: valve B: fire control valve
Similar to	"Entity A is similar to Entity B" means that A and B is similar in function or concept	A: plug B: socket
Contain	"Entity A contains Entity B" means that B is a physical component of A	A: air conditioning B: air compressor

The ResCNN model took 200 epochs of training to achieve stable performance. Subsequently, a test process was performed to evaluate the prediction accuracy of the model, and the results are shown in Table 4. The general accuracy of the model is around 70%, which is relatively acceptable. However, the performance of the model is much poorer than that of the NER model, probably because here the training data set needs to be manually labeled and is thus much smaller. After the program finished, a manual inspection process was also carried out to improve the quality of the relationships. As a result, a total of 9439 relations were discovered.

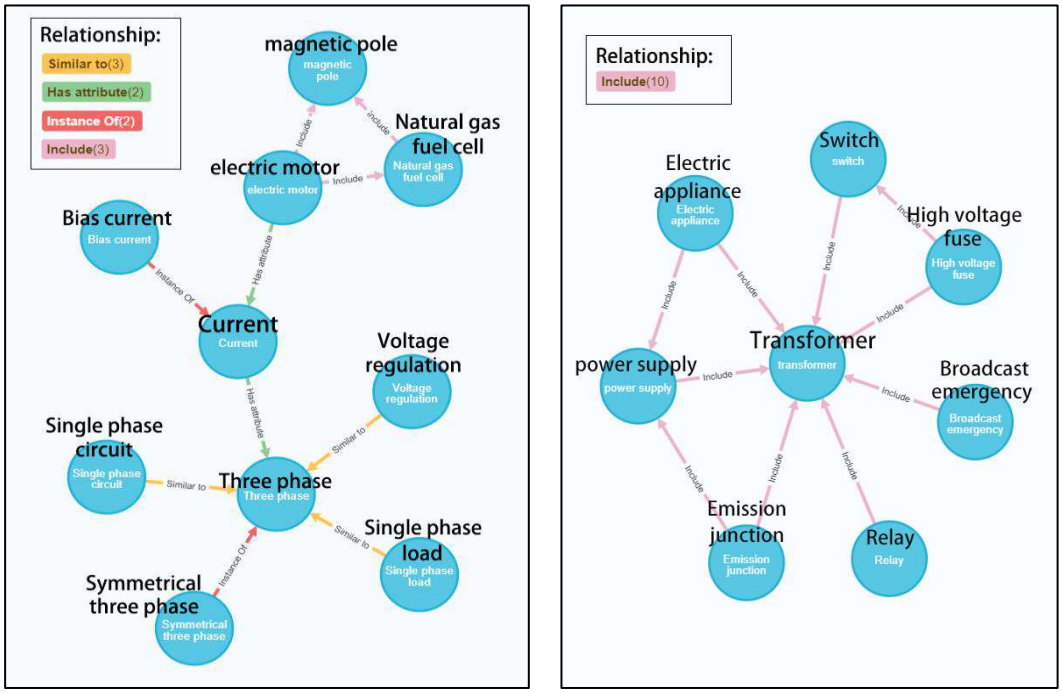
Table 4: Training results of the ResCNN model

Relationship categories	Precision	Recall	F1 Score
No relationship	57.43	75.35	64.67
Instance of	58.90	89.87	70.69
Similar to	87.00	84.37	85.17
Contain	79.07	51.37	61.80
Total	70.60	75.24	70.58

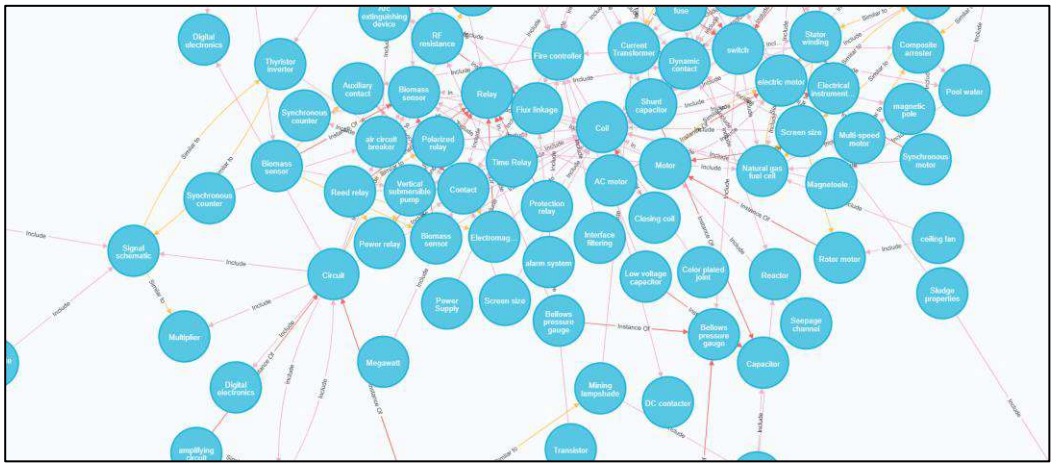
After the knowledge acquisition process, the entities and relationships were merged together to construct a knowledge graph. In the proposed approach, a graph database Neo4j (Neo4j, 2019) was applied to store and visualize this knowledge. The graph database records data in the form of nodes and edges, which enables faster modification and query speed for graph-based data than traditional relational databases. In this graph database, the MEP entities were applied as nodes in the graph, and their relationships were taken as edges. A part of the constructed knowledge graph is shown in Fig. 4a. Compared with plain texts or tables, the knowledge graph can present MEP knowledge more systematically, concisely and intuitively. A more complex part of the graph is shown in Fig. 4c, illustrating the knowledge is connected together to form a network structure.

Based on the graph database, the knowledge query function can be further implemented. For example, one might be interested in what MEP facilities contain a transformer as a component. The graph database can quickly return results to him from tens of thousands of knowledge with just one command, as shown in Fig. 4b. This query is hard to accomplish for text-based data even with the help

of the Internet.



(a). part of the Knowledge graph *(b). Search result of "Transformer"*



(c). a more complex part of the Knowledge Graph

Figure 4: Visualization of the MEP Knowledge graph (Translated into English)

5. Discussion and Conclusion

Although only simple knowledge-graph-based functions such as visualization and node search were achieved in this paper, the role of knowledge graph goes far beyond this. Proved by the famous search engine Google, knowledge graph can sort out the domain knowledge without the help of experts and become the basis of ontology construction. Moreover, the knowledge graph can be used as databases of various AI applications such as reasoning and dialog system.

Though the method to establish the MEP knowledge graph is tested to be feasible in this research,

it should be pointed out that the proposed knowledge graph is far from perfect and requires further researches and works. The scale of the graph needs to be expanded as the data source is not comprehensive and the amount of data is insufficient. At the same time, the knowledge extraction algorithm needs to be modified to improve its performance. Knowledge fusion and disambiguation also need to be processed to improve the quality of the graph.

As a conclusion, a MEP knowledge graph which is a systematic summary of the fragmented texts in the network was constructed in this research. To achieve the automatic MEP knowledge acquisition, a diverse range of MEP text documents were obtained from the network, and multiple NLP techniques, including word segmentation, entity extraction and relationship discovery, were applied to discover entities and relationships from the text. This knowledge was then organized in the form of a network to represent a knowledge graph. A simple application of the knowledge graph was presented, and possible applications were further discussed. The knowledge graph constructed in this paper is expected to contribute to the promotion of AI technology in the MEP industry.

Acknowledgements

This research was supported by the National Key R&D Program of China (Grant No. 2017YFC0704200), the National Natural Science Foundation (No. 51778336), and the Tsinghua University – Glodon Joint Research Center for Building Information Modelling.

References

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american, 284*(5), 28-37.

Bouzidi, K. R., Fies, B., Faron-Zucker, C., Zarli, A., & Thanh, N. L. (2012). Semantic web approach to ease regulation compliance checking in construction industry. *Future Internet, 4*(3), 830-851.

Fabian, M., Gjergji, K., & Gerhard, W. (2007). *Yago: A core of semantic knowledge unifying wordnet and wikipedia*. Paper presented at the 16th International World Wide Web Conference, WWW.

Fxsjy. (2019). Jieba Chinese text segmentation. Retrieved from https://github.com/fxsjy/jieba

Hu, Z.-Z., Tian, P.-L., Li, S.-W., & Zhang, J.-P. (2018). BIM-based integrated delivery technologies for intelligent MEP management in the operation and maintenance phase. *Advances in Engineering Software, 115*, 1-16.

Hu, Z.-Z., Zhang, J.-P., Yu, F.-Q., Tian, P.-L., & Xiang, X.-S. (2016). Construction and facility management of large MEP projects using a multi-Scale building information model. *Advances in Engineering Software, 100*, 215-230.

Huang, Y., & Wang, W. Y. (2017). *Deep Residual Learning for Weakly-Supervised Relation Extraction*. Paper presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jiang, S., Wang, N., & Wu, J. (2018). Combining BIM and Ontology to Facilitate Intelligent Green Building Evaluation. *Journal of Computing in Civil Engineering, 32*(5), 04018039.

Korman, T. M., Fischer, M. A., & Tatum, C. B. (2003). Knowledge and reasoning for MEP coordination. *Journal of Construction Engineering and Management, 129*(6), 627-634.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., . . . Auer, S. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic web, 6*(2), 167-195.

Neo4j. (2019). Neo4j Graph Platform – The Leader in Graph Databases. Retrieved from https://neo4j.com/

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web, 8*(3), 489-508.

Pauwels, P., Zhang, S., & Lee, Y.-C. (2017). Semantic web technologies in AEC industry: A literature overview. *Automation in Construction, 73*, 145-165.

Sawant, U., Garg, S., Chakrabarti, S., & Ramakrishnan, G. (2019). Neural architecture for question answering using a knowledge graph and web corpus. *Information Retrieval Journal*, 1-26.

Turk, Ž. (2006). Construction informatics: Definition and ontology. *Advanced engineering informatics, 20*(2), 187-199.

Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences, 112*, 112-120.

Wang, H., Miao, X., & Yang, P. (2018). *Design and Implementation of Personal Health Record Systems Based on Knowledge Graph*. Paper presented at the 2018 9th International Conference on Information Technology in Medicine and Education (ITME).

Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., & Guo, M. (2018). *RippleNet: Propagating user preferences on the knowledge graph for recommender systems.* Paper presented at the Proceedings of the 27th ACM International Conference on Information and Knowledge Management.

Wang, L., & Leite, F. (2016). Formalized knowledge representation for spatial conflict coordination of mechanical, electrical and plumbing (MEP) systems in new building projects. *Automation in Construction, 64*, 20-26.

Wu, Z., & Tseng, G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science, 44*(9), 532-542.

Zhou, Z. (2010). *Manufacturing Intelligence for Industrial Engineering: Methods for System Self-Organization, Learning, and Adaptation*: IGI Global.