Towards Accurate Search for E-Commerce in Steel Industry: A Knowledge-Graph-Based Approach

Maojian Chen[1,2,3], Hailun Shen[4], Ziyang Huang[4], Xiong Luo[1,2,3(✉)], and Junluo Yin[1,2,3]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
xluo@ustb.edu.cn
[2] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
[3] Shunde Graduate School, University of Science and Technology Beijing, Foshan, Guangdong 528399, China
[4] Ouyeel Co., Ltd., Shanghai 201999, China

Abstract. Mature artificial intelligence (AI) makes human life more and more convenient. However, in some application fields, it is impossible to achieve the satisfactory results only depending on the traditional AI algorithm. Specifically, in order to avoid the limitations of traditional searching strategies in e-commerce field related to steel, such as the inability to analyzing long technical sentences, we propose a collaborative decision making method in this field, through the combination of deep learning algorithms and expert systems. Firstly, we construct a knowledge graph (KG) on the basis of steel commodity data and expert database, and then train a model to accurately extract steel entities from long technical sentences, while using an advanced bidirectional encoder representation from transformers (BERT), a bidirectional long short-term memory (Bi-LSTM), and a conditional random field (CRF) approach. Finally, we develop an intelligent searching system for e-commence in steel industry, with the help of the designed KG and entity extraction model, while improving the searching performance and user experience in such system.

Keywords: Steel E-commerce · Knowledge graph (KG) · Entity extraction · Bidirectional encoder representation from transformers (BERT)

1 Introduction

Living in the era of big data, more and more data are generated on the Internet, and more choices are available for people, which makes it difficult for executives to make decisions [1]. In order to find the best strategy, the collaborative decision making (CDM) process is developed. It is a flexible process considering every aspect to obtain the best

H. Gao et al. (Eds.): CollaborateCom 2020, LNICST 349, pp. 3–18, 2021.
https://doi.org/10.1007/978-3-030-67537-0_1

benefit [2]. During the last decade, CDM is applied in many fields [3–8]. However, as far as the steel which is used as an essential material in our life, there is little research on CDM applied to the steel e-commerce field.

Among the e-commerce field in the steel industry, the search system is directly facing users to service. It emphasizes efficiency which means that the assigned tasks are completed with the shortest time while considering stability. Meanwhile, the accuracy of searching results and the efficiency of search system are largely influencing user experience, and user's satisfaction with the system reflects the pursuit of the steel e-commerce field.

Traditional steel search engines retrieve information on the Internet through keywords, and return relevant webpages containing strings to users. With the increasing complexity of business data, this retrieval method cannot accurately satisfy users' diverse demands, and it will greatly affect user' experience. For example, because some non-standard steel commodity information is filled in when uploading commodity information to the trading platform by traders, these existing search engines cannot retrieve and analyze complex information with multiple attributes well. Meanwhile, there are some informal vocabularies in user's inquiries, which may lead to an unsatisfactory ability to analyze steel daily inquiries. Then, they do not support the retrieval of the commodities' nickname, nor the retrieval of steel daily inquiry. In order to avoid the limitations of traditional steel search engines, new search engines on the basis of knowledge graph (KG) have attracted extensive attention from relevant researchers [9–12]. Hence, the introduction of KG will provide a new way for search engines of e-commerce trading platforms in steel industry.

Currently, there are some works on the metal and materials related to the steel industry. For example, it has been verified that constructing KG of steel enterprise operation and maintenance domain can obtain dispersed and heterogeneous information in a consistent way, while simplifying the process of obtaining information and improving the efficiency of obtaining information for engineers [13]. Then, by using DBpedia and Wikipedia, a method was developed to construct metallic materials KG [14]. A KG was constructed, where it included 115 materials properties and 69 relationships. This KG can represent arbitrarily complex property and relationships [15].

With the development of steel industry, it has accumulated amount of steel data forming expert databases. However, the above methods are using accurate algorithms or convenient tools to achieve the sort of steel data, rather than utilizing expert databases to improve the user experience and user satisfaction for the steel e-commerce.

Hence, in consideration of the above background, and to further effectively handle those difficulties, we combine steel commodity data and expert databases to construct a KG in steel industry, and standardize commodity information. Furthermore, through the use of bidirectional encoder representation from transformers (BERT) [16], bidirectional long short-term memory (Bi-LSTM) [17], and conditional random field (CRF) approach [18], we further optimize and improve the performance of existing search engine, so as to make the match between users' demands and commodities more accurate, make steel trading simpler and more effective, and improve user's experience.

The rest of this paper is organized as follows. Section 2 will introduce the related work, including KG, entity extraction and entity alignment model. In Sect. 3, we detail

the method on the search system of steel e-commerce platform using KG. Then, we will introduce the experiments, including data sets, experimental results, and some discussions in Sect. 4. Finally, the conclusion and future work will be summarized in Sect. 5.

2 Related Work

In this section, we will introduce some key technologies in related to our method.

2.1 Knowledge Graph (KG)

In 2012, KG was proposed by Google, in an effort to optimize existing search engines [19]. Different from traditional search engines, the KG-based search engines can comprehend user's intentions from the semantic level and extract complicated information better, thus effectively improving the search performance.

Essentially, KG is a semantic network. Its nodes represent entities or concepts, and edges represent various semantic relationships between entities and concepts [20, 21]. With the full use of visual technology, KG can not only describe the knowledge resources and carriers, but also analyze and the relations between them [22]. In recent years, KG becomes one of the basic technologies in intelligent services, such as semantic search, intelligent question answering, and decision support [23].

The core technologies of KG involve knowledge extraction, knowledge representation, knowledge fusion, and some others. Knowledge extraction is used to extract entities, concepts, attributes, and relationships from various data sources. Knowledge representation means that the semantic information of entities can be expressed with dense low dimensional vectors, and then entities, relationships, and complicated semantic associations among entities can be efficiently calculated in a low dimensional space. Knowledge fusion includes semantic computing and data integration, which are to eliminate contradictions and ambiguities [24, 25].

Currently, there are many KG-based applications in different fields. In respect of search engines, whether Google, Bing, or Baidu, they all implement intelligent search on the basis of KG. For instance, if users want to retrieve information about "Yao Ming" in Bing, it will return a knowledge card of "Yao Ming" instead of some webpages containing the string "Yao Ming". The knowledge card shows Yao's date of birth, height and weight, the names of his spouse and children. This returned method can improve retrieval efficiencies greatly.

2.2 Unstructured Daily Queries Analysis

When a customer provides a daily query, it needs to be parsed to understand his requirements. Daily queries can be divided into three categories, i.e., a mixture of multiple keywords, Chinese and English mixed keywords, and sentences. If the query is just a number of keywords, it is relatively simple to parse. If it is a sentence, it needs to be deeply analyzed, which involves semantic analysis, entity extraction, and entity alignment. In this paper, we focus on the daily queries analysis with a sentence.

Entity extraction technique was proposed in 1996 [26], and it is automatically extracting various entities from text or sentence, like persons, time, locations, money [27]. Now, entity extraction is a basic technology of Chinese text analysis, and there are many machine learning methods applied to it [28, 29], such as hidden markov models (HMM), support vector machines (SVM).

The entity extracted from the daily queries may be a nickname, that is not consistent with the standard entity. Thus, these extracted entities cannot be used directly in other processes, and they need to be aligned.

There are some methods on entity alignment, most of which calculate the similarity between two entities to decide whether they are the same entity. IF-IDF algorithm counts the frequency of each character in an entity, and then transforms those frequencies into vectors, lastly calculates the similarity of the two vectors through cosine distance [30–32]. Word2Vec is to map each entity into a low-dimensional space to form a shorter vector, and then calculates the similarity of the two vectors through cosine distance [33, 34]. This method takes the semantic of the entity into account, and it is better than TF-IDF. In addition, there are string-based, corpus-based, knowledge-based, and other methods that can be used to calculate the similarity of two entities [35–39].

3 Methodology

In this section, we will introduce how to construct a KG in the steel industry with CDM process, and then develop a search system of e-commerce field. The whole framework is shown in Fig. 1.

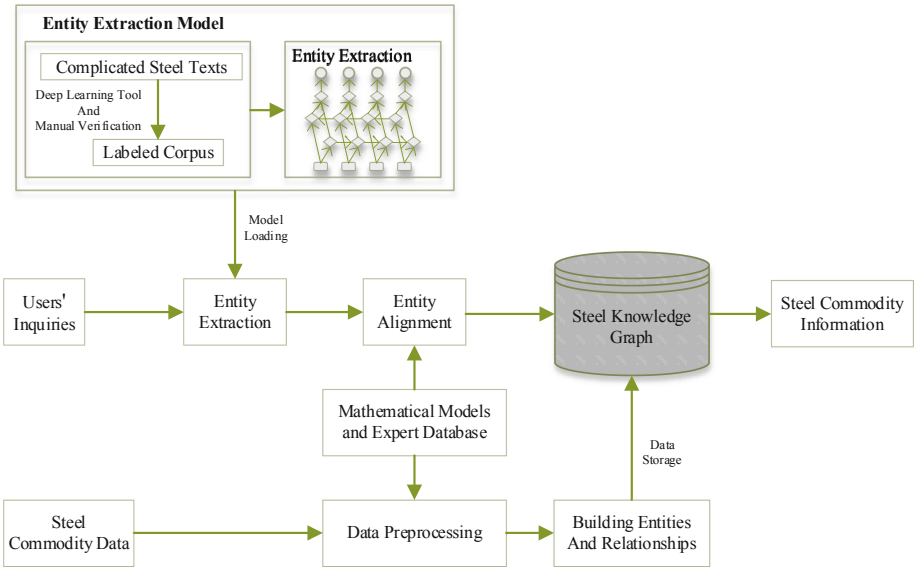

Fig. 1. The framework of establishing a KG-based search system for e-commerce platform.

Firstly, we construct a steel KG. In this processing, since the data of steel commodity are structured, we need to use mathematical models and expert database to extract important data. Then, according to the attributes of steel commodities, we establish entities and relations between entities, and those entities and relations will be stored in graph database. Secondly, in order to improve the computing performance, we use a third-party tool to label the unstructured steel data, and then manually verify them. Thirdly, to accurately comprehend users' intentions, some advanced deep learning algorithms are used to train these data to obtain an entity extraction model for the steel data. Fourthly, the trained entity extraction model is used to extract users' inquiries for getting entities and align entities. Finally, the aligned entities are matched with the entities in the graph database, and the corresponding steel commodity information is returned to users.

3.1 Preprocessing for the Steel Commodity Data

Firstly, we collect a lot of the steel commodity data from the Internet. Secondly, we clean and filter those data through mathematical models and expert database. Thirdly, we select some important data as entities, and construct relationships between those entities. Then, we store those data into graph database to construct a KG.

3.2 Preprocessing for the Steel Inquiry Data

Our purpose is to parse users' daily inquiries, thus we use the historical users' daily inquiry data as dataset. However, these data are unlabeled and unstructured. Therefore, we need to label those data first.

Among these steel data, there are many types of entities, such as numeral, technical term, name of a person, place name, and many others. Hence, we segment those data and tag them with their part-of-speech (POS). Then, the BMEO principle is used to further label the position of each character in the entity. Here, "B" means that the character is the first word in the entity, "M" represents that the character is the internal word in the entity, "E" shows that the character is the last word in the entity, and "O" expresses that the word is not an entity. Finally, we divide these data into three parts, including training set, validation set, and test set.

3.3 Entity Extraction Model

In this section, we will introduce how to use unstructured steel texts to train an entity extraction model.

Word Embedding with BERT. The data of the steel industry are text, while the input of training model must be numeral, therefore, we need to normalize those data. Here, we use BERT model to generate word embedding. Generally, the dimension of word embedding is less than the number of words.

Entity Extraction via Bi-LSTM+CRF Model. Bi-LSTM is a special recurrent neural network (RNN), which is mainly used to deal with the issue of vanishing gradient and exploding gradient during long sequence training.

The main idea of Bi-LSTM is to retain historical information and remember current information, by introducing a gate mechanism and controlling the degree of each unit. In so doing, it can retain important features and discard unimportant features. Totally, Bi-LSTM performs better in longer sequences than common RNN.

In this paper, the output of Bi-LSTM is the scores for each tag of the character. For example, if there are 7 tags, and the output of Bi-LSTM is shown as Fig. 2. For input w_0, the score of tag "B-Person" is 1.624, the score of tag "M-Person" is 0.819, the score of tag "E-Person" is 0.203, the score of tag "B-Organization" is 0.765, the score of tag "M-Organization" is 0.050, the score of tag "E-Organization" is 0.101, and the score of tag "O" is 0.035. It can select the tag with the highest score in each character as the result. But, in some cases, such as it described in Fig. 3, the result of two characters' tags in the entity "w_0w_1" tagged with "M-Organization" and "M-Person" is wrong clearly, since it is impossible for two middle characters of different categories to be adjacent.

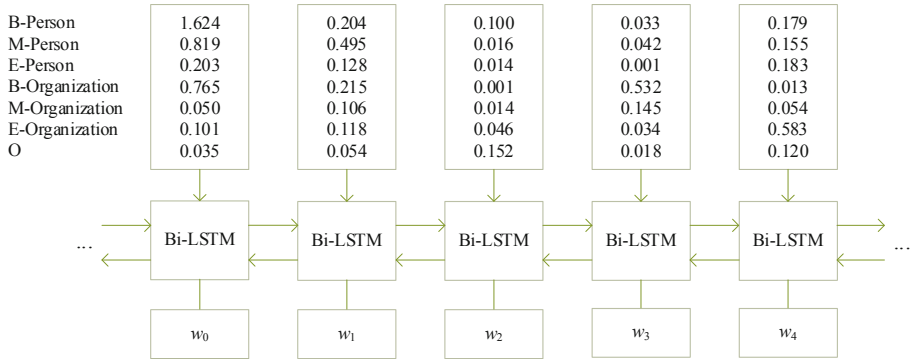

Fig. 2. An example of the output in Bi-LSTM.

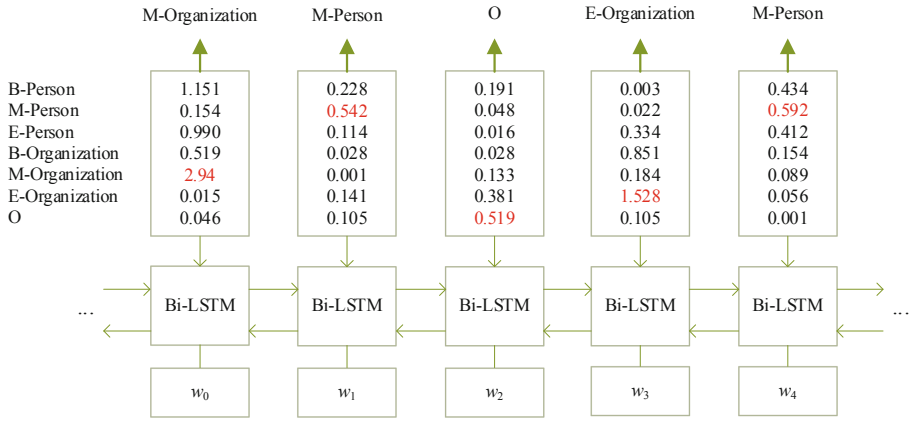

Fig. 3. An example of the result of Bi-LSTM.

CRF can automatically learn some restrictions during training to ensure that the predicted tags are valid. There are some restrictions showing as follows.

(1) The first character's tag in a sentence should start only with "B-" or "O", rather than others.
(2) For the tag result as "B-tag1 M-tag2 M-tag3 E-…", here tag1, tag2, tag3 should be the same tag. For example, "B-Person E-Person" is valid, and "B-Person M-Organization" is invalid.
(3) "O-tag" is illegal. Entities should start with "B-", not "M-" or "E-".

When these constraints are learned, the number of invalid tags in the final results is reduced dramatically. The structure of Bi-LSTM+CRF model is shown in Fig. 4.

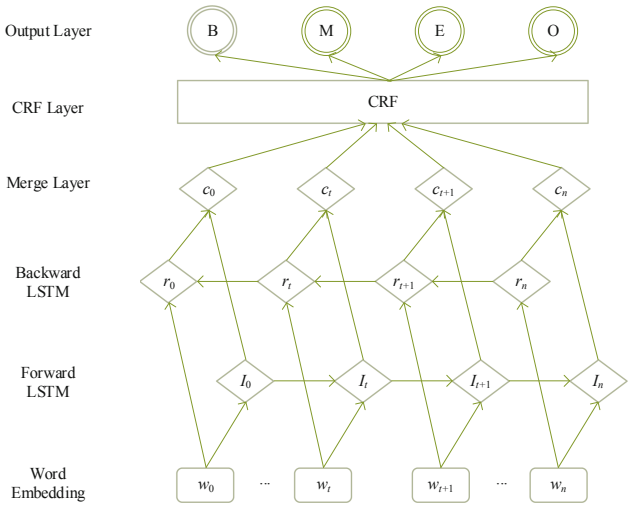

Fig. 4. The structure of Bi-LSTM+CRF model.

3.4 The KG-Based Search System

After obtaining the entity extraction model and KG, we accordingly develop a search system in relation to the steel e-commerce field. The whole process is as follows.

Firstly, when the customer inputs a daily inquiry, the system extracts entities from the inquiry according to the trained model. Because the extracted entities contain many professional terms in the steel industry, the expert database in the steel industry is used to align entities. Then, the entities obtained in the previous step are matched with the KG. Furthermore, we extract the IDs of the commodity, and they contain all the entities. Finally, the steel commodity information corresponding to the commodity IDs are returned to the customer.

In so doing, a KG-based search system for the steel e-commerce can be constructed.

4 Experiments

Different from the common industries, the data of the steel industry are more complex and professional. Therefore, it is necessary to combine expert knowledge to preprocess the

data and construct KG. Generally, there are many graph databases, such as ArangoDB, Neo4j, and JanusGraph. In our experiments, we select Neo4j 4.1.1 as our graph database. Meanwhile, our experiments are conducted in the Python 3.6.8 environment running on Ubuntu 18.04.1.

4.1 Datasets and Data Preprocessing

Here, two datasets are used in our experiments, i.e., the steel commodity data, and the daily inquiry data in steel industry.

There are 20,366 steel commodity data, including more than 30 entity categories. Those data are available through public ways on the Internet. Considering that there are many dirty data in dataset, we use mathematical models and expert database to clean up those data, such as filtering attributes, handling invalid data and missing data. In this process, the mathematical models are used to extract useful and important attribute value from original steel commodity data, such as defining the rules used to extract the weight of objective. Through the combination of the mathematical models and the expert database, we are able to extract defects, surface treatment, coating type, and other attribute values from original steel commodity data.

The sample of the steel commodity data is shown in Table 1. From the attribute value of "zero spangle Z80 cr free", we can see that the "Z80" is coating weight, rather than coating type, so it needs to be cleaned and preprocessed with mathematical models and expert database.

Table 1. The samples of the original steel commodity data.

Grade	Specification	Place	Resource ID	Coating type
SGC570	0.65*1240*C	Tang Steel	1398950769	Zero spangle Z80 cr free
HC420LAD+Z	1.65*1405*C	Ben Steel	1432699197	
DX53D+Z	0.8*1650*C	Shou Gang	1433283775	Light oiling

It is easy to obtain the steel commodity data, however, it is relatively difficult to collect unstructured steel inquiry data, since the inquiry data exist in the communication tools between customers and suppliers, such as email, social software. Through our cooperation partner, we got 2,600 daily inquiry data, which are from customers' historical purchase information. Therefore, all data we used in this paper are real and reasonable. After cleaning low available and repeated data, there remain 1,217 useful inquiry data.

The sample of the unstructured steel inquiry data is shown in Table 2. In this table, the expressions of customer's demands are informal. For example, in first inquiry data, "DC54D+Z" means that the grade of steel plate is "DC54D", and coating type is "Zinc". Similarity, for the "SCGA270D-45" in the third data, its grade is "DX52D", and coating type is "Zinc-Fe". After translating Table 2 from Chinese into English, the updated version is shown in Table 3.

Table 2. The samples of the unstructured steel inquiry data.

Index	非结构化的钢铁询单数据
1	求购:谁家有 DC54D+Z 2.5*66*1250的,我要一张
2	2.0的电镀锌SECCN5的样板,求购一块
3	求购SCGA270D-45,要2吨
4	上海地区求购 0.5*1000环保钝化无花

Table 3. The updated version of Table 2 after translating it from Chinese into English.

Index	Unstructured steel inquiry data
1	Who has DC54D+Z and 2.5*66*1250 steel plate, I want to buy one
2	Purchase SECCN5 steel plate with 2.0 electric galvanizing
3	Purchase 2 tons SCGA270D-45 steel plate
4	Purchase 0.5*1000 environmental passivation and zero zinc coil in Shanghai

4.2 Experiments and Result Analysis

In addition to data preprocessing mentioned above, there are other three parts in our experiments.

Constructing KG Based on Commodity Data in Steel Industry. After preprocessing the data, we select five important attributes of the steel commodity data as entities used in Neo4j, with the help of experts in steel industry. They are "Grade", "Coating Type", "Coating Weight", "Surface Structure", and "Surface Treatment". Moreover, we define the commodity ID, which is the central entity to connect those five types of entities. Then, the KG of steel industry is accordingly constructed. The storage architecture of steel commodity data in KG is shown as Fig. 5, and nodes with different colors represent different categories entities. For example, the red entity means "Grade", the green entity represents "Coating Type", and the orange entity is "Surface Structure". After translating Fig. 5 from Chinese into English, the updated version is shown in Fig. 6. There are 1,057 entities and 1,778 relationships stored in the Neo4j database, and a part of KG of steel commodity data is shown in Fig. 7. From this figure, we can easily find some information, e.g., for the Commodity ID "SP0009", whose surface structure is "Zero Spangle", and its coating weight is "40/40". After translating Fig. 7 from Chinese into English, the updated version is shown in Fig. 8.

Word Embedding. We use the BERT model to achieve word embedding of steel inquiry data. Here, in consideration of the characteristics of steel commodity data, we convert each word to a 128-dimensional vector. The sample of word embedding in steel industry is shown in Table 4.

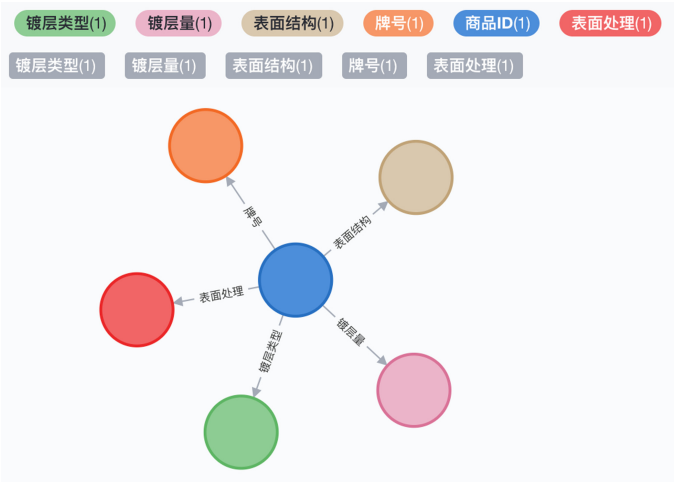

Fig. 5. The architecture of steel commodity data in KG.

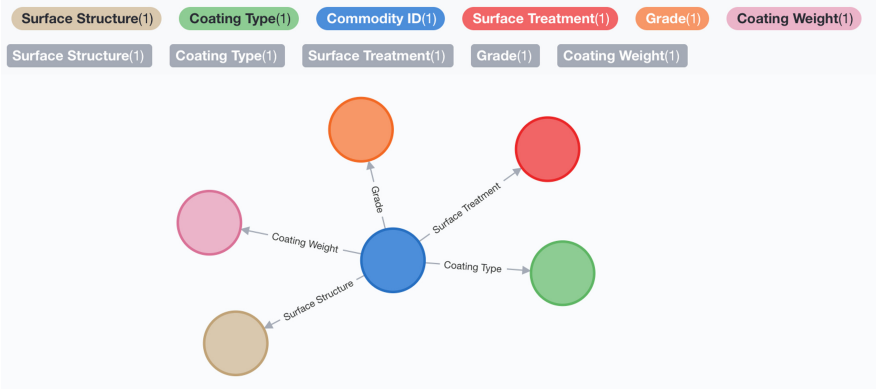

Fig. 6. The updated version of Fig. 5 after translating it from Chinese into English.

Entity Extraction Model. For the steel user's inquiry data, we use Jieba tool for Chinese words segmentation and POS tagging firstly, and then manually verify the results. In our experiments, we have defined eight entity-tag categories, which are "Grade", "Place", "Specification", "Thickness", "Weight", "Surface Structure", "Surface Treatment", and "Species". Additionally, we use BMEO principle to label the position of each character in the entity. Through this method, we can get 25 different POS tags. The example of above result is shown in Fig. 9. Firstly, the inquiry data "Purchase hot-dip zinc coating coil, cr free, zero spangles" has segmented and tagged simply with the Jieba tool and manual intervention. Here, each character in the entity whose tag belonging to the above eight tags will be further tagged with the BMEO principle.

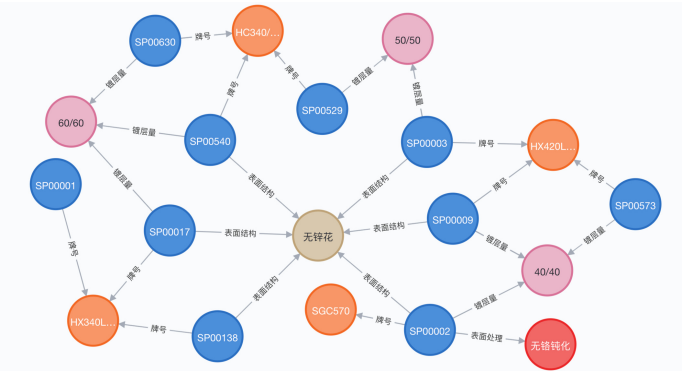

Fig. 7. A part of KG in displaying steel commodity data.

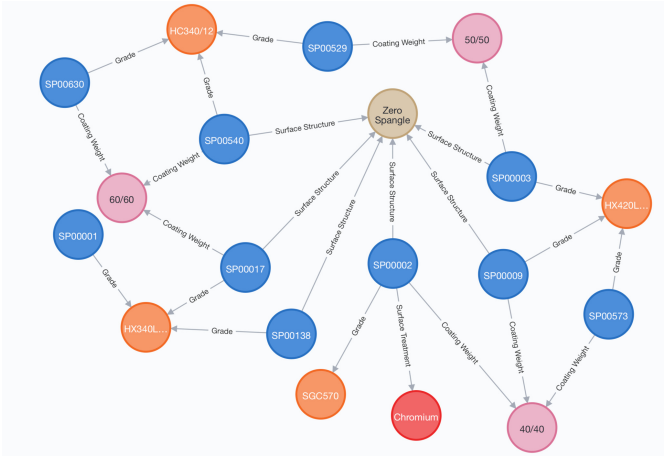

Fig. 8. The updated version of Fig. 7 after translating it from Chinese into English.

Table 4. The samples of word embedding obtained by BERT in the steel industry.

Chinese character	Word embedding
热	−0.28838387 0.5150681 …−0.4132588
镀	−0.44589975 0.41127107 …0.04480578
锌	−0.17859079 0.52611357 …0.04148377
C	−0.39107212 0.18615262 …−0.41394806
料	−0.25390878 0.8519357 …−0.42595372

Note: “热镀锌C料” appeared in this table means “GI Coil Grade DX51D”.

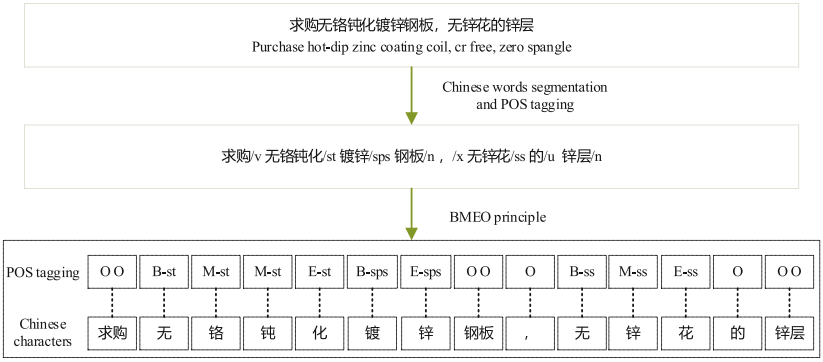

Fig. 9. The result of Chinese words segmentation, POS tagging and BMEO principle.

Through the use of BERT and Bi-LSTM+CRF model, the entity extraction model for the steel industry is trained. When inputting a sentence, the model can extract steel entities accurately.

Performance Comparison. In order to evaluate the performance of the BERT+Bi-LSTM+CRF algorithm in the steel industry, we compare it with Word2Vec+Bi-LSTM+CRF and BERT+CRF algorithms, and the result is shown in Table 5. Here, through many experiments, we set the training epoch is 100, the size of the training batch is 8. From Table 5, by comparing three metrics, precision, recall and F1, we find that BERT+Bi-LSTM+CRF performs better than other two algorithms.

Table 5. The performance comparison using BERT+Bi-LSTM+CRF, Word2Vec+Bi-LSTM+CRF and BERT+CRF in the steel industry.

Method	Precision of test data	Recall of test data	F1 of test data
Word2Vec+Bi-LSTM+CRF	87.31%	64.31%	0.74
BERT+CRF	87.50%	90.25%	0.89
BERT+Bi-LSTM+CRF	**89.71%**	**91.83%**	**0.91**

The KG-Based Search System for Steel E-Commerce. Based on the KG and entity extraction model, we develop a search system used in the steel e-commerce field. When a customer inputs a sentence as Fig. 10, there are 7 entities that can be extracted. After aligning entity with mathematical models and expert database, there are only 4 important entities related to the steel industry. By matching the 4 entities with the KG and comprehensively considering the expert system, we can find 43 related steel commodity IDs. Finally, we return the corresponding information of steel commodity data to the customer through these IDs. After translating Fig. 10 from Chinese into English, the updated version is shown in Fig. 11.

Fig. 10. The result of KG-based search system for steel industry.

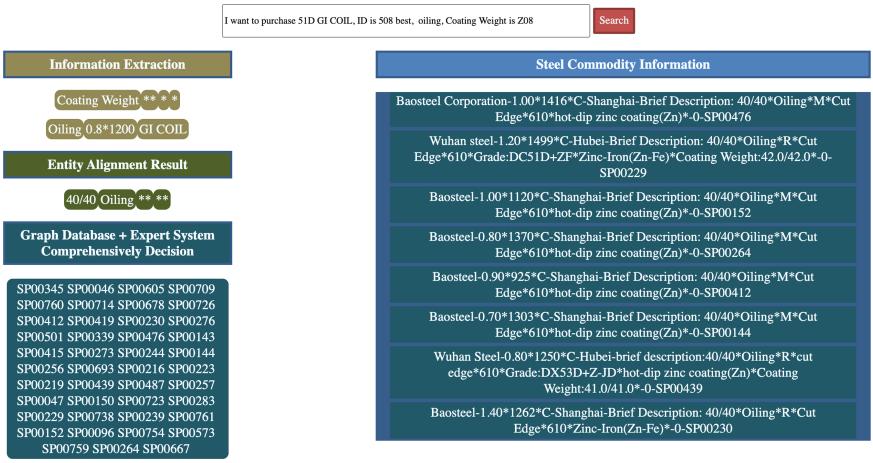

Fig. 11. The updated version of Fig. 10 after translating it from Chinese into English.

5 Conclusion

In this paper, through the incorporation of the CDM process into the steel e-commerce field, we construct a KG for the steel industry, train an entity extraction model which can accurately extract entities about steel from sentences, and develop a search system for the steel e-commerce, to serve the purpose of improving the system performance and optimizing user experience. Firstly, we collect steel commodity data from the Internet, and clean those data with some mathematical models and expert database in the steel industry to filter some important attributes. Secondly, these attributes as entities are stored in Neo4j graph database. Thirdly, Jieba tool, Chinese words segmentation, POS tagging, and BMEO principle are used for user inquiry data. Fourthly, BERT, Bi-LSTM,

and CRF models are applied to train these data to obtain an entity extraction model that can extract entities well for the steel industry. Finally, on the basis of the KG and entity extraction model, we construct a steel search system for steel e-commerce to improve searching performance and user experience.

In the future, for expanding the scale of KG in the steel industry, we will collect more and more steel commodity data from the Internet, select more attributes as entities, and optimize the storage architecture of the data in the KG. Specifically, we will also collect more daily inquiry data about steel to improve the accuracy of the entity extraction model. In addition, we will further train a new model that can automatically label the unstructured steel data through BERT+Bi-LSTM+CRF.

Acknowledgment. This work was supported in part by the National Key R&D Program of China under Grant 2016YFC0600510, in part by the Beijing Natural Science Foundation under Grant 19L2029, in part by the Beijing Intelligent Logistics System Collaborative Innovation Center under Grant BILSCIC-2019KF-08, in part by the Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB, under Grant BK19BF006, and in part by the Fundamental Research Funds for the University of Science and Technology Beijing under Grant FRF-BD-19-012A.

References

1. Benali, M., Ghomari, A.R., Zemmouchi-Ghomari, L., Lazar, M.: Crowdsourcing-enabled crisis collaborative decision making. Int. J. e-Collaboration **16**(3), 49–72 (2020)
2. Campbell, C., Roth, W., Jornet, A.: Collaborative design decision-making as social process. Eur. J. Eng. Educ. **44**(3), 294–311 (2019)
3. Chun, S., Shulman, S., Sandoval, R., et al.: Government 2.0: making connections between citizens, data and government. Inf. Polity **15**(1), 1–9 (2010)
4. Kapucu, N., Arslan, T., Demiroz, F.: Collaborative emergency management and national emergency management network. Disaster Prev. Manage. **19**(4), 452–468 (2010)
5. Lyndon, M., Angela, C.: Collaboration in health care. J. Med. Imaging Radiat. Sci. **48**(2), 207–216 (2017)
6. Hsiao, F., Zeiser, S., Nuss, D., Hatschek, K.: Developing effective academic accommodations in higher education: a collaborative decision-making process. Int. J. Music Educ. **36**(2), 244–258 (2018)
7. MacDonald, A., Clarke, A., Huang, L.: Multi-stakeholder partnerships for sustainability: designing decision-making processes for partnership capacity. J. Bus. Ethics **160**(2), 409–426 (2019)
8. Peng, P., Li, Y., Zhou, L.: Research on interactive collaborative decision-making method of equipment support task planning. In: 5th International Conference on Computer and Communication Systems, pp. 533–537. IEEE, Shanghai (2020)
9. Ebisu, T., Ichise, R.: Generalized translation-based embedding of knowledge graph. IEEE Trans. Knowl. Data Eng. **32**(5), 941–951 (2020)
10. Liu, Q., Li, Y., Duan, H., Liu, Y., Qin, Z.: Knowledge graph construction techniques. J. Comput. Res. Dev. **53**(3), 582–600 (2016)
11. Van Luijt, B., Verhagen, M.: Bringing semantic knowledge graph technology to your data. IEEE Softw. **37**(2), 89–94 (2020)
12. Zhou, J., Sun, X., Yu, X., Bian, X.: Knowledge graph and data application: intelligent recommendation. Telecommun. Sci. **35**(8), 165–172 (2019)

13. Jin, G., Lü, F., Xiang, Z.: Enterprise information integration based on knowledge graph and semantic web technology. J. SE Univ. (Nat. Sci. Edn.) **44**(2), 250–255 (2014)
14. Zhang, X., Liu, X., Li, X., Pan, D.: MMKG: An approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia. Comput. Phys. Commun. **211**(2), 98–112 (2017)
15. Mrdjenovich, D., Horton, M., Montoya, J.H., Legaspi, C.M., Dwaraknath, S., Tshitoyan, V., Jain, A., Persson, K.A.: Propnet: a knowledge graph for materials science. Matter **2**(2), 464–480 (2020)
16. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186. ACL, Minneapolis, MN, USA (2018)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
18. Lafferty, J., Mccallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: 18th International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2002)
19. Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., Van De Walle, R.: Adding realtime coverage to the Google knowledge graph. In: ISWC Posters and Demonstrations Track, pp. 65–68. CEUR-WS, Boston (2012)
20. He, P.: Counter Cyber Attacks by Semantic Networks: Emerging Trends in ICT Security. Morgan Kaufmann, Boston (2014)
21. Rahman, A.: Knowledge representation: a semantic network approach. Handbook of Research on Computational Intelligence Applications in Bioinformatics (2016)
22. Zhang, Y., Liu, X., Bai, X., Yin, J.: Collaborative research on intelligence perception and characterization of search engines. J. Beijing Inf. Sci. Technol. Univ. **34**(6), 19–24 (2019)
23. Huang, H., Yu, J., Liao, X., Xi, Y.: Review on knowledge graphs. Comput. Syst. Appl. **28**(6), 1–12 (2019)
24. Paulheim, H., Cimiano, P.: Knowledge graph refinement: a survey of approaches and evaluation methods. Seman. Web **8**(3), 489–508 (2017)
25. Zhang, Y., Dai, H., Kozareva, Z., Smola, A.J., Song, L.: Variational reasoning for question answering with knowledge graph. In: 32nd AAAI Conference on Artificial Intelligence, pp. 6069–6076. AAAI Press, New Orleans (2018)
26. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: 16th International Conference on Computational Linguistics, pp. 466–471, Copenhagen (1996)
27. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: 27th International Conference on Computational Linguistics, pp. 2145–2158. ACL, New Mexico (2018)
28. Goyal, A., Gupta, V., Kumar, M.: Recent named entity recognition and classification techniques: a systematic review. Comput. Sci. Rev. **29**, 21–43 (2018)
29. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. arXiv: 1812.09449 (2018)
30. Tata, S., Patel, J.: Estimating the selectivity of tf-idf based cosine similarity predicates. ACM SIGMOD Rec. **36**(4), 7–12 (2007)
31. Albitar, S., Fournier, S., Espinasse, B.: An effective TF/IDF-Based text-to-text semantic similarity measure for text classification. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014. LNCS, vol. 8786, pp. 105–114. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11749-2_8
32. De Boom, C., Van Canneyt, S., Bohez, S., Demeester, T., Dhoedt, B.: Learning semantic similarity for very short texts. In: IEEE International Conference on Data Mining Workshop, Atlantic City, NJ, USA, pp. 1229–1234. IEEE, New York (2015)

33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: 26th International Conference on Neural Information Processing Systems, Nevada, pp. 3111–3119 (2013)
34. Wu, C., Wang, B.: Extracting topics based on Word2Vec and improved Jaccard similarity coefficient. In: IEEE Second International Conference on Data Science in Cyberspace, Shenzhen, China, pp. 389–397. IEEE, New York (2017)
35. Kenter, T., Rijke, M.: Short text similarity with word embeddings. In: 24th ACM International on Conference on Information and Knowledge Management, pp. 1411–1420. ACM (2015)
36. Huang, G., Guo, C., Kusner, M., Sun, Y., Weinberger, K.Q., Sha, F.: Supervised word mover's distance. In: 30th Conference on Neural Information Processing Systems. Neural Information Processing Systems Foundation, Barcelona, Spain, pp. 4862–4870 (2016)
37. Blanco, E., Moldovan, D.: A semantic logic-based approach to determine textual similarity. IEEE/ACM Trans. Audio Speech Lang. Processing **23**(4), 683–693 (2015)
38. Smarandache, F., Colhon, M., Vlăduţescu, Ş., Negrea, X.: Word-level neutrosophic sentiment similarity. Appl. Soft Comput. **80**, 167–176 (2019)
39. Lee, Y., Ke, H., Yen, T., Huang, H., Chen, H.: Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. J. Assoc. Inf. Sci. Technol. **71**(6), 657–670 (2020)