Construction of knowledge graph of forest musk deer based on BiLSTM-CRF model and DPA method

Chuqiao Yang*, Haiyan Wang*, Jingyan Sai, Peiwei Zhang, Mengyao Yan
School of Information, Beijing Forestry University, Beijing 100083, China
yangchuqiao@bjfu.edu.cn
wanghaiyan2008@bjfu.edu.cn

Abstract. Forest musk deer is a national key protected species, but due to overhunting and habitat destruction, wild musk deer has disappeared in large areas. Constructing a knowledge graph in the field of forest musk deer can standardize the relevant knowledge in the field of forest musk deer and lay a good foundation for wildlife protection and ecological construction. In order to construct the knowledge map of the forest musk deer, the entities were further subdivided first, and the BiLSTM-CRF model was used to identify the entities of the BIO-labeled data. After analysis, it was found that the model used in the experiment performed better. Secondly, the Dependency Parsing Analysis method was used to identify the relationship, and the relationship between the entities was smoothly extracted. Finally, the Neo4j graph database was used to realize the storage and visualization of the knowledge graph of the forest musk deer. This research transforms the unstructured text in the field of forest musk deer into structured data. Through the systematic sorting of relevant knowledge in the field of forest musk deer, it can provide wildlife protection, artificial breeding of forest musk deer, intelligent knowledge services and other aspects some new technical means in the field of forest musk deer.

1. Introduction

China is the country with the most abundant species and resources of musk deer, and the production of musk deer resources accounts for more than 90% of the world[1]. Historically, Chinese musk deer is widely distributed, but since the 20th century, due to excessive hunting and habitat destruction, our country's wild musk deer resources have been severely damaged. Although a series of measures have been taken in recent years to protect forest musk deer and attempt to restore wild musk deer populations, due to the wide distribution of the remaining wild musk deer populations and low population density, the natural recovery process is very slow[2]. In order to solve the contradiction between the supply and demand of natural musk and the reintroduction of artificially cultivated forest musk deer, China successively started experiments on artificial musk breeding in China in 1958[3]. After decades of groping for aquaculture, some progress has been made in artificial breeding of forest musk deer in China. However, due to the lack of theoretical guidance, population quality degradation, and artificial breeding of musk deer diseases, the development of the forest musk deer breeding industry is still relatively slow[4].

With the rapid development of the Internet era, people have more convenient ways to obtain knowledge. The systematic arrangement of forest musk deer related knowledge with the help of knowledge graph technology is of great significance to related research work in the field of forest musk deer. The existing knowledge maps of vertical fields abroad mainly include IMDB in the field of
film and television, MusicBrains in the field of music[5], GeoNames in the field of geography[6], DBLife in the academic field, and UniProtKB in the biological field. There are also related studies in China. For example, Jia Lirong et al. [7] did a knowledge map construction of Chinese medicine; Wang Renwu et al. [8] explored the construction of a Chinese business knowledge map based on deep learning and graph databases. However, in wild animals In the field, due to the wide distribution of data sources, the lack of data structure, and the lack of relevant research and reports, application research based on knowledge graphs is very rare.

The knowledge graph theory was originally proposed by C. Hoede and F. N. Stokman et al. It uses graphs to show the internal structure and development of the discipline[9]. Sowa proposed the basic theory of concept maps, including a large number of relationship types, such as equivalence, order, causality, similarity, etc. [10]. Su[11] proposed a semantic search engine framework. This framework will first create a domain-oriented ontology, and then use crawlers to crawl information. The information extraction team will propose named entities from the crawled information, and then classify them into specific topics. Under the body. Based on the above-mentioned knowledge graph theory, until 2012, Google Company proposed the concept of knowledge graph for the first time[12] and successfully applied it to semantic search, which improved the quality of search results.

This research will construct a knowledge map of the forest musk deer, use the BiLSTM-CRF model to perform entity recognition on the data annotated by BIO, use Dependency Parsing Analysis to extract relations, and finally use the Neo4j graph database to realize the storage and visualization of the forest musk deer knowledge map. The knowledge map in the field of forest musk deer constructed in this study will provide a new perspective for wildlife protection and ecological construction.

2. Experimental Method and Model Design
The experimental method for constructing the knowledge map of the forest musk deer in this study is shown in Figure 1. First, the crawler technology was used to obtain the experimental source data, and the data is cleaned and annotated; then the BiLSTM-CRF model is used to identify the data and introduce the subdivision entity. Improve the accuracy of entity recognition; then use Dependency Parsing Analysis to extract the relationship between entities, and finally use the Neo4j graph database to fully express the entities and relationships. Through data collection, data processing, knowledge graph display and other links, the disordered knowledge in the forest musk deer field can be displayed in a regular manner and provide better knowledge services for the forest musk deer.

![Fig.1 Experimental method for constructing forest musk deer knowledge graph](image-url)
2.1. Entity acquisition

Accurately obtaining entities from the corpus is one of the most basic and important tasks in the construction of the knowledge graph. This study will use cyclic neural networks to train unstructured texts in the field of forest musk deer, and finally obtain the forest musk deer from the target corpus. The knowledge entity of the domain. Isozaki et al. [13] used SVM in entity recognition tasks and obtained good results; Hu Wenbo et al. [14] aimed at the insufficiency of traditional CRF that can not well recognize complex geographical names and organization names, and proposed a cascaded CRF identification method, and large-scale tests have been carried out, and good results have been achieved. Collobert et al. [15] proposed a window-based deep neural network model, which automatically learns abstract features from input sentences, and trains model parameters through a backward propagation algorithm, which has achieved good results beyond traditional algorithms.

The Bi-directional Long Short-Term Memory (BiLSTM) network model can obtain context information and splice the two kinds of information together to obtain the final calculation result. Since the BiLSTM model cannot directly use the predicted labels, the prediction for each text sequence is independent, which may cause the final prediction result to not comply with the labeling rules. For example, using BIO sequence labeling, The correct label sequence in a sentence is "B-Loc I-Loc P", but the predicted result is "I-Loc I-Loc O", which does not meet the labeling rules. Conditional Random Field (CRF) is introduced to solve the problem of non-compliance with rules. The addition of the conditional random field model allows the model to learn some rules and conditions, thereby adding some constraints to the BiLSTM model to ensure that the predicted label output by the BiLSTM model meets the requirements.

The BiLSTM-CRF model used in this paper is mainly composed of three layers: Look-up Layer, Bidirectional Long Short-term Memory Network Layer (BiLSTM Layer) and Conditional Random Field Layer (CRF Layer). Its structure is shown in Figure 2.

![Fig. 2 Composition of BiLSTM-CRF model](image)

2.2. Relation extraction

Relationship extraction is another basic and important step in the process of knowledge graph construction. The quality and accuracy of relationship extraction are of great significance to the construction of knowledge graph. Entity relationship extraction needs to identify the entity information contained in the text and the relationship between entities. Liu Shaoyu et al. [16] used the SVM classifier to determine the region of the sample, and used the K-nearest neighbor algorithm to classify the samples in the fuzzy region, and finally obtained better experimental results on the relation extraction data set of the SemEval-2010 evaluation task; Kang Qi [17] proposed an automatic domain knowledge acquisition method based on Bootstrapping, which can automatically extract domain entities, their corresponding attributes and attribute values from web pages with only a small amount of artificially defined initial seeds.
In this study, we mainly use Dependency Parsing Analysis (DPA) to obtain the relationship between knowledge entities in the forest musk. DPA believes that the core verb in a sentence is the central component that governs other words, and the verb is not affected by any other words. Domination, therefore, all the dominated words depend on the core verbs in a certain Chinese relationship. The annotation relationship of dependency parsing is shown in Table 1.

HanLP can be used for word segmentation and DPA. The HanLP tool is a java toolkit composed of a series of models and algorithms. It is an open source Chinese natural language processing software. Its goal is to popularize the application of natural language processing in the production environment, with perfect functions, efficient performance, clear architecture, corpus up-to-date and customizable.

| Relationship type                  | Annotation relationship | Relationship type          | Annotation relationship |
|-----------------------------------|-------------------------|----------------------------|-------------------------|
| Subject-predicate relation        | SBV                     | Verb-object relation       | VOB                     |
| Inter-object relation             | IOB                     | Prepositional object       | FOB                     |
| Concurrent language               | DBL                     | Modifier-head relation     | ATT                     |
| the structure of Adverbial Modifier | ADV                     | Verb-complement relationship | CMP                    |
| Coordinative relation             | COO                     | Prepositional relation     | POB                     |
| Left attachment relation          | LAD                     | Right attachment relation  | RAD                     |
| Absolute construction             | ACD                     | Core relation              | HED                     |

3. Experiment and Result Analysis

The data sources of the experiment include encyclopedia websites such as Baidu encyclopedia and Sogou encyclopedia, multimedia interactive platforms such as WeChat official account, Weibo and Zhihu, news websites such as the People's Daily, Xinhua, and local news websites, as well as academic websites such as CNKI and WanFang Database. Get articles about forest musk deer from these sites by using the corpus collectors. A total of 1523 related articles are collected, and invalid articles were removed. A total of 1393 articles are valid.

3.1. Entity extraction

This experiment uses the BIO labeling method to tag the corpus. The BIO labeling method is to tag each entity that needs to be recognized in the following way: "BN", "IN", "O" (where N represents a certain knowledge entity), "BN" means the first word of the knowledge entity, "IN" means other parts of the knowledge entity other than the first word, and "O" means the part of a sentence that is not the knowledge entity. A total of 1415 sentences containing knowledge entities are selected from the 1393 articles crawled, and the required 9 types of entities are marked (as shown in Table 2).

| Entity type       | label  |
|-------------------|--------|
| forest musk deer  | FMDEER |
| musk              | SX     |
| musky action      | USE    |
| conservation Area | BHQ    |
| mountains, forests| MOUNT  |
| another name      | NAME   |
| place             | LOC    |
The training architecture for this experiment is TensorFlow 1.9.0, the batch is set to 20, the learning rate is set to 0.001, in the output of the BiLSTM layer, the dropout value is set to 0.5, the number of iterations is 40, and the Adam algorithm is used for optimization. The feature selection is the character vector corresponding to each word of the input corpus, and the character vector is obtained by using the character vector trained on the Internet.

In model work, the setting of the number of hidden layer nodes has an important impact on the prediction results. Too few nodes may make the model learning insufficient, and too many nodes will cause the model structure to be complex and performance degradation. In this experiment, the number of hidden layer nodes is set to 150, so that a 300-dimensional output vector can be learned.

The experiment uses the above configuration for training, a total of 416 entities are labeled, 414 entities are obtained, and 399 entities are correctly identified. The experimental results obtain after calculation based on the above evaluation mechanism are: accuracy (P) is 96.38%, the recall rate (R) is 95.91%, and the F1 value is 96.14. The entity recognition experiment results of each type of entity are shown in Table 3. From the experimental results in the table, the model used in the experiment has good effect and can accurately identify most required entities.

| Type of   | Precision | Recall  | F1-measure |
|----------|-----------|---------|------------|
| FMDEER   | 100%      | 99.02%  | 99.51%     |
| SX       | 100%      | 100%    | 100%       |
| USE      | 100%      | 100%    | 100%       |
| BHQ      | 87.5%     | 93.33%  | 90.32%     |
| MOUNT    | 100%      | 100%    | 100%       |
| NAME     | 100%      | 100%    | 100%       |
| LOC      | 98.26%    | 95.36%  | 96.79%     |
| ANIMAL   | 100%      | 100%    | 100%       |
| YZC      | 80%       | 100%    | 88.89%     |

3.2. Relation Extraction
The HanLP tool is used for DPA. Firstly, load the previously obtained entity as a custom dictionary into the HanLP tool. After importing the custom dictionary, the result of Dependency Parsing Analysis through HanLP can directly obtained. Figure 3 showed the result of DPA for the corpus, which is “Sichuan Baishuihe National Nature Reserve lies in Pengzhou city in the southeast of the Longmen Mountains. The vegetation types of the reserve include coniferous forest, coniferous and broad-leaved mixed forest, deciduous broad-leaved forest, evergreen broad-leaved forest, shrub, meadow and sparse vegetation of rocky beach.”
Fig. 3 Analysis results of DPA

In the result of the DPA, the first column outputs the position label of the word in the sentence, followed by the corresponding word and relationship type, and the last column is the position of the pointing word and the pointing word. The relationship between words can be clearly obtained through the DPA. Because of this feature, the results obtained from DPA can be combined with the method of machine learning to form an extraction mode in the subsequent relationship extraction process. According to the results of DPA, the relationship and orientation between words can be clearly obtained. Therefore, if the grammatical relationship between words is known, the program can be written to extract triples. Figure 4 shows that the result is the extraction result of the subject and predicate object in the sentence shown in Figure 3. After manual screening, a total of 859 relationships are obtained, which can be divided into 7 categories: location, interior, breeding, function, production, title and other relationships.

4. Neo4j Knowledge Graph Visualization

Graph database is a kind of non-relational database (NoSQL), which is suitable for expressing the data of graph structure. Its basic meaning is to store and query data with the data structure of "graph". A graph contains two data types: nodes and relationships, both of which can contain attributes in the form of key-value pairs, and nodes are connected by relationships to form a relational network structure. Neo4j is a main representative of graph databases, and it comes with a set of query language called Cypher, which is similar to the SQL language and can operate on the database. Due to the large number of nodes and relationships, the visualization interface of Neo4j only loads 300 nodes and relationships by default, click to expand more relationships.

The experimental results can also be displayed using an independent research and development system, using Eclipse IDE as a development tool, J2EE as a development basis, using maven junction
and SSM framework, using tomcat7 as the server, neo4j as the database for development, and using Echarts’ force-directed graph to visualize the knowledge graph. Figure 5 is the query result of forest musk deer in Sichuan Province about the protected area, and Figure 6 is the query result of the forest musk deer name. Through the above examples, we can find that the visual knowledge graph can effectively show the relationship between single or multiple entities.
5. Conclusion

Based on relevant data in the field of forest musk deer, after further subdividing the entities, the BiLSTM-CRF model is used to perform entity recognition on the BIO-labeled data set. 9 types of entities are successfully identified, and a total of 585 knowledge entities are obtained. Relations were extracted using Dependency Parsing Analysis combined with manual screening, and a total of 860 relations were obtained. Neo4j was used to store the final knowledge graph, which promoted the in-depth development of the forest musk deer field from the perspective of information technology.

Although the experimental results of this study have achieved relatively good results, there are still shortcomings, such as the effect of dependency parsing still needs to be improved, and the ability to automatically build knowledge graph needs to be improved. It is expected that breakthroughs and solutions will be made in the following research. The knowledge graph in the field of forest musk deer is a comprehensive summary of the knowledge in the field of forest musk deer. Its construction can bring certain scientific guidance to the artificial breeding industry of forest musk deer, and can provide a new perspective for scientific research in the field of forest musk deer. It can be used in wildlife protection, intelligent knowledge services and other aspects to provide new technical means. With the emergence and development of new technologies such as Artificial Intelligence, Big Data, Cloud Computing, and the Internet of Things, the related applications of knowledge graph in the field of wildlife protection will develop rapidly, which presumably will bring outstanding contributions to the ecological construction of the entire industry.

Acknowledgments

This work was financially supported by the National Training Program of Innovation and Entrepreneurship for Undergraduates of Beijing Forestry University (G201910022061).

References

[1] Helin Sheng. Current Situation of Musk deer Resources in China and Rescue measures[J]. Wild Animals.1996(03):10-12. (in Chinese)
[2] Linhai Li. The Status and Developing strategies of Musk deer Farming Industry in China[D]. Beijing Forestry University, 2012. (in Chinese)
[3] Ji Hua. Reintroduction Planning of Forest Musk deer and Habitat selection Research in Release site[D]. Northwest A&F University, 2016. (in Chinese)
[4] Chengli Zheng, Yonghua Cai, Jianmin Wang, Chengxu Wang, Lei Zhou, Dayong Feng, Ying Yang. Investigation Report on the Current Situation of Forest Musk Breeding in Sichuan Province[J]. Heilongjiang Animal Science and Veterinary Medicine, 2020(03):136-139. (in Chinese)
[5] Schartz A. Musicbrainz: A semantic web service[J]. IEEE Intelligent Systems, 2002, 17(1):76-77.
[6] Goodwin J, Dolbear C, Hart G. Geographical linked data: The administrative geography of great britain on the semantic web[J]. Transactions in GIS, 2008.
[7] Lirong Jia, Jing Liu, Tong Yu, Yan Dong, Ling Zhu, Bo Gao, Lihong Liu. Construction of TCM Knowledge Graph [J]. Journal of Medical Informatics, 2015, 36(08):51-53+59. (in Chinese)
[8] Renwu Wang, Yi Yuan, Xuping Yuan. Exploratory research on Constructing Chinese Business Knowledge graph based on Deep learning and Graph database [J]. Library and Information, 2016(01):110-117. (in Chinese)
[9] Yue Chen. The Development history of the Scientific knowledge map [J]. Studies in Science of Science. 2008, 26(3). (in Chinese)
[10] J. F. Sowa. Conceptual Structures: Information Processing in Mind and Machine. Web Semantics: Science, Services and Agents on the World Wide Web, 1984, 113-124.
[11] S. Su, D. Wu, T. Yu, J. Z. Li, S. Wang, Z. Q. Li. A Framework of Service Oriented Semantic Search. International Conference on Computational Problem-Solving, 2001, 499-504.
[12] Singhal A. Introducing the knowledge graph: things, not strings[J]. Official google Blog, 2012.
[13] Isozaki H, Kazawa. Efficient Support Vector Classifiers for Named Entity Recognition[C].
International Conference on Computational Linguistics. DBLP, 2002:390-396.

[14] Wenbo Hu, Yuncheng Du, Xueqiang Lv, Shuicai Shi. Chinese named Entity recognition based on Multi-layer conditional Random field [J]. Computer Engineering and Applications, 2009, 45(1):163-165. (in Chinese)

[15] Collobert R, Weston J, Bottow L, et al. Natural Language Processing (Almost) from Scratch [J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[16] Shaoyu Liu, Bicheng Li, Zhigang Guo and so on. Summary of Research on Entity Relationship Extraction [J]. Journal of Information Engineering University, 2016, 17(5):514-547. (in Chinese)

[17] Qi Kang. Research on Automatic Extraction of Domain Knowledge Based on Bootstrapping [D]. Shanong University, 2012. (in Chinese)