Earthquake Knowledge Graph Constructing Based on Social Intercourse Using BiLSTM-CRF

Xiaoqing Sun1,3, Li Qi2, Hai Sun4, Wenbo Li1, Cheng Zhong5, Yiliang Huang6 and Ping Wang7

1 Institute of Technology Innovation, Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei, China
2 The Third Research Institute of the Ministry of Public Security, Shanghai, China
3 The Department of Automation, University of Science and Technology of China, Hefei, China
4 China Telecom Shanghai Ideal Information Industry (Group) Co., Ltd, Shanghai, China
5 Beijing Aerospace Flight Control Center, Beijing, China
6 Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China
7 Department of Airborne Remote Sensing, National Disaster Reduction Center of China, Beijing, China
Email: ylhuang@iim.ac.cn

Abstract. Nowadays, the earthquake has become a very serious topic. The earthquake-related information always appears first in social media. Constructing an earthquake knowledge graph can help dealing with earthquake news social media text data. This paper proposed the BiLSTM-CRF (Bi-directional Long Short Term Memory-Conditional Random Field) model to construct earthquake knowledge graph using news text data. The BiLSTM-CRF model identified the entities and then writes entities and their types to the table so that the entities and relationships between entities can be extracted for earthquake. The entities and relationships between entities have been combined with the form of RDF (Resource Description Framework) to construct earthquake knowledge graph on Neo4j database.

1. Introduction
In recent years, earthquakes have occurred more and more frequently. According to China Earthquake Network, only over the past years, there have been 125 earthquakes over 6 magnitudes in the world [1]. More importantly, frequent earthquakes have caused a large amount of casualties [2, 3], buildings damage [4-6], economic losses [7-9] and other negative effects [7-12]. Nowadays, the earthquake-related information always appears first in social media [13-16]. Earthquake is a public security incident, so it’s important to deal with the earthquake-related social media information. Knowledge graph [20] can be selected to organize and analyze these historical data and assist in inferring new information from social media data [17-20]. Knowledge graph is a structured connection network between various entities through certain relationships such as DBpedia, YAGO [20], etc. Knowledge graph can not only be easily understood by people through visualization, but also make it easier for computers to process various data and information [20]. Knowledge graph is similar to the human network, which consists of lots of relationships between people. The smallest unit is people and the relationship between them.

In order to use the knowledge graph in a certain scene, it is necessary to construct a specific
knowledge graph [20]. NER (Named Entity Recognition) [21] as well as relation extraction [21] is the basic step in knowledge graph building. Some researchers have used a dictionary-based approach [21] to achieve entities information extraction, such as Proux [22]. This method compares the unknown information with the dictionary library and uses the matched information as the final identification information [20], but this approach requires a very complete dictionary which is currently unattainable. Some researchers have used a rule-based approach to achieve this purpose, such as Rau [23], who have constructed an entity extraction prototype system that automatically extracts company names from text. This method requires manual construction of rules and information extraction based on rules, but there is also some trouble that it need a lot of manpower and resources, and has a very poor scalability. In order to get rid of the above shortcomings, machine learning technologies, such as CRF (Conditional Random Fields) [21], SVM (Support Vector Machine) [21] and deep learning, have been used to extract entity information. Machine learning has been a learning method that uses computers to simulate human behavior [21]. Deep learning is a branch of machine learning [21]. More importantly, deep learning mimics the transmission of the neural network in the human brain [21]. Constructing knowledge graph with deep learning mainly consists of three parts. NLP (Natural Language Processing) [25] can be processed using deep learning [26], which involves word segmentation, entity recognition and relationship extraction [21]. NLP uses the tagging database that labels the information to train and learn unknown information through the trained model. However, these techniques require to handle the extraction of entities and relationships using machine learning, which takes a lot of time, the entity types in news text data can often be equivalent to relationships, so relationship extraction that using machine learning can often be replaced by direct extraction.

In this paper, the BiLSTM-CRF (Bi-directional Long Short Term Memory-Conditional Random Field) [27] model was proposed to construct earthquake knowledge graph based on social intercourse. BiLSTM-CRF is one kind of deep learning models. What’s more, the proposed BiLSTM-CRF can be used to directly extract the entities and the entity types as relationships [28, 29].

2. Methods

2.1. Main Part of Construction of Knowledge Graph
This knowledge graph was mainly divided into three parts: named entity recognition [21], data sorting, entity relationship extraction(see Figure 1). Then, following part will introduce each part of the content.

![Figure 1. The Main Part of Construction of Knowledge Graph](image)

2.1.1. Named Entity Recognition. The process of identifying the required entities from the news text data. Entity recognition is the first and most important step in the construction of knowledge graph. After recognizing the entities, the following work can be done. In this work, the BiLSTM-CRF [27] model was to recognize entities in earthquake news text data. This model is the commonly used model for named entity recognition in the deep learning [21].

2.1.2. Data Sorting. The follow-up to the named entity recognition, used to write the identified entity and its entity type to the table. The entity type was written as the relationship to the first row as the attribute of the relational table. The identified entity was used as the tuple of the relational table. In this article, the entities that used for classification was written into the first column in order to extract
important entity types as the relationship between the entities under the entity type and the classification entities.

2.1.3. Entity Relationship Extraction. Firstly, the first column of the records was extracted as the first element of the triple. Secondly, the other entities that corresponding to it were extracted as the third element of the triple. Finally, the attributes were written as the second element of the relationship to the triple.

2.2. BiLSTM-CRF Model

2.2.1. LSTM. LSTM (Long Short Term Memory) is a special RNN (Recurrent Neural Network) [30]. It is used to make up for the defects that normal recurrent neural networks cannot remember for a long time [30]. This model has a relatively common RNN for short-term memory, and a library for storing long-term information [30]. The RNN part also has some unique parts: the forgetting gate, the input gate, and the output gate for controlling the network (see Figure 2) [30].

![Figure 2. The Structure of LSTM](image)

Input gate:

\[ i_t = \sigma \left(W_i \cdot [h_{t-1}, x_t] + b_i \right) \]  \hspace{1cm} (1)

\[ C'_t = \tanh \left(W_c \cdot [h_{t-1}, x_t] + b_c \right) \]  \hspace{1cm} (2)

Forgetting gate:

\[ f_t = \sigma \left(W_f \cdot [h_{t-1}, x_t] + b_f \right) \]  \hspace{1cm} (3)

Update status:

\[ C_t = f_t \cdot C_{t-1} + i_t \cdot C'_t \]  \hspace{1cm} (4)

Output gate:

\[ o_t = \sigma \left(W_o \cdot [h_{t-1}, x_t] + b_o \right) \]  \hspace{1cm} (5)

\[ h_t = o_t \cdot \tanh \left(C_t \right) \]  \hspace{1cm} (6)
The input of the model is the output $h_{t-1}$ of the previous moment, the state vector $C_{t-1}$ of the previous moment and the input $x_t$ of the moment. The output is $h_t$ and the state vector $C_t$ at that moment.

In this paper, the BiLSTM (Bi-directional Long Short Term Memory) was proposed (see Figure 3) to predict corresponding label probabilities belonging to earthquake news entities after being trained with the related annotated data set.

2.2.2. CRF. CRF (Conditional Random Field) [31] is a discriminant model that used to calculate the probability under a certain condition. Its essence is to satisfy features saved in the memory base to help the next train.

$$P(Y_i | X, Y_w, w! = v) = P(Y_i | X, Y_w, w = v)$$

(7)

Figure 4. Gaining the Label Prediction Probability under the Condition of Word Boundary Condition

It is the probability of meeting an event under certain condition [31]. This probability is constrained by a certain condition and achieves the expected effect more accurately. In this paper, the CRF model was used to obtain the label prediction probability based on work boundary condition after several most possible label was chose (see Figure 4).

2.2.3. BiLSTM-CRF. The BiLSTM-CRF (Bi-directional Long Short Term Memory-Conditional Random Field) model is the combination of Bidirectional LSTM model and CRF model (see Figure 5) [27]. The Bidirectional LSTM model is in the lower layer, and it is used to predict the likelihood of the label of the inputs [27]. The CRF model is in the upper layer and it is used to calculate the conditional probability prediction labels [27]. Compared with the single BiLSTM model, this model can contact the corpus context of the world to improve the accuracy of the prediction.
2.3. Neo4j Database
This part mainly introduced a graph database—Neo4j database [32]. This database has been relatively mature and currently widely used. This database can not only store a large amount of structured data, but also visualize it to complete the functions required in this article. Its interface has been very simple: the command line has been provided at the top, and the command line can be used for cypher operation to complete the database storage function. Programming languages such as java, python, etc. can also complete the docking through API (Application Programming Interface). The left side can visually view the current build’s type name and nodes, relationships, etc. as well as some instance scripts and basic query messages (see Figure 6).

The Neo4j database has some common cypher operation commands [32]: (1) ‘CREATE’ command, which is used to create the required node, relationship as well as attributes. (2) ‘RETURN’ command, which can be used to return the specified content. (3) ‘MATCH’ command [4], which can be used to acquire data by comparing with data stored in the database. In this paper, we used the ‘CREATE’ command as an example to operate this database (see Figure 7):

```
CREATE (n: disaster {disaster_NAME: 'earthquake'});
```
This command created a node that belongs to the earthquake type and named it ‘earthquake’.

---

**Figure 6. Neo4j Database Interface**

---

**Figure 5. BiLSTM-CRF Model**
3. Experiments and Results

3.1. Preparation
Some preparatory work was required before building the news-based earthquake knowledge graph. To begin with, it was necessary to obtain news text data which related to earthquake. In this paper, we selected the 2003-2015 World Seismic News Text data that had been posted on the news page. But the data on the news data source was imperfect and needed to be supplemented later. Then, an annotated data set in textual form related to the type of data we extracted was also considerable. Only after training the annotation dataset, we employed the training model to recognize entities from the data source after the training model obtained. What’s more, converting the annotated dataset to a dictionary form was also required. After preparing the above three parts, we completed the preparation part for earthquake knowledge graph constructing based on social intercourse.

3.2. Training and Recognizing.
This part was aimed at recognize entities and it was divided into training part, test part and recognizing part. The training set and the test set were extracted in a 3:1 ratio in the annotated data set dictionary. The BiLSTM-CRF model was used for training, and the accuracy, recall and F value were obtained respectively during training and testing (see Figure 7). The training model was then used to recognize the entities in the data source to obtain the entities and their types.

| Set   | Accuracy | Recall | F value |
|-------|----------|--------|---------|
| Train | 0.808    | 0.738  | 0.772   |
| Test  | 0.690    | 0.610  | 0.647   |

According to the above table, the F value obtained by training and testing was low, and the effect of the recognition was poor, so further adjustment and improvement of its parameters were needed.

3.3. Data Sorting
Writing the identified entities to the table was the main part of this work. We did the following before writing the entities: determined whether the earthquake was an earthquake and the intensity level
based on the recognized magnitude, and replaced the magnitude with the intensity level and replaced the entity type with the ‘belong event’. Then join whether the earthquake was an earthquake and write its entity type as ‘Whether earthquake’. If no magnitude was detected, we set the intensity level to ‘nan’. Wrote the entity types to the table column, and wrote the entity records to the table row, which was recognized in each earthquake event. Note that the ‘belong event’ type needed to be written to the first column.

3.4. Extracting and Building Triples

This part extracted the entities and relationships written in the table. The relationships of the part were the entity types to which the entity belonged. In order to represent it as the triplet form of entity-relationship-entity, the first column entities were the first element of the triple. The other entities of this row were the last element of the triple, and the entity types were equivalent to the relationship between the first element and the last element and then as the second element of the triple. This resulting triplet was the knowledge graph represented by the RDF (Resource Description Framework) form.

3.5. Graph Storage and Visualization

This section mainly introduced the storage and visualization of the news-based earthquake knowledge graph from the social intercourse using the proposed BiLSTM-CRF. In order to apply the graph in the subsequent work or to reason and supplement the graph, it was necessary to save the constructed knowledge graph, and the location where the knowledge graph was saved was the database. And visualization can display the knowledge graph in the form of RDF to judge the performance of the constructed knowledge graph to a certain extent. One visual way was to store it with the graph database so that the stored graph can be visualized by the graph database in the form of graph structure. Then came out and completed desired performance. The graph database we selected was the Neo4j database, which is a relatively mature graph database system. By using this database, the storage and visualization functions can be well implemented. After storing the knowledge graph in the form of RDF into the Neo4j database, the visualization of earthquake knowledge graph from social intercourse was obtained (see Figure 8).

![Figure 8. News-based Earthquake Knowledge Graph](image)

3.6. Discussion

This paper mainly discussed the construction of the earthquake knowledge graph from the social intercourse using the BiLSTM-CRF. There were some improvements over other existing work:

1. For the particularity of news text data, the entity types were equivalent to the relationship between entities. The existing works based on deep learning has extracted entities and relationships simultaneously using deep learning, while deep learning requires long events. This way saved the time required for relationship extraction based on deep learning.

2. Obtain entity types such as ‘belong event’ and ‘whether earthquake’ intensity by ‘magnitude’ entity type. According to common sense, the ‘for’ loop was used to judge whether the earthquake under a certain magnitude was a strong earthquake or a weak earthquake. Whether the event was an
earthquake by whether or not recognize the magnitude. In this way, the entity types and the number of entities can be expanded to some extent.

But there were still some shortcomings:

(3) News text data as a data source was not complete. We selected some earthquake news from 2003-2015 as a data source for identification, which was incomplete and did not capture all earthquake news data for this time period and other time periods.

(4) The accuracy of entity recognition was not high enough. In this paper, the $F$ value of training and test was only 0.772 and 0.647, and this number can be improved.

(5) The construction of this knowledge graph was still relatively rough. We will divide this relation table into multiple relational tables to improve this problem.

4. Acknowledgments
This paper is financially supported by the National Key Research and Development Program of China (No. 2018YFB1004600), the National Natural Science Foundation of China (No. 41871302, 61773360, 41701594).

5. References
[1] http://www.ceic.ac.cn/speedsearch?time=10
[2] Wyss M and Chamlagain D 2019 Estimated casualties in possible future earthquakes south and west of the M7.8 Gorkha earthquake of 2015 ACTA GEOPHYS vol 67 pp 423-429
[3] Zhao B 2016 April 2015 Nepal earthquake: observations and reflections NAT HAZARDS vol 80 pp 1405-1410
[4] Brunner D, Lemoine G and Bruzzone L 2010 Earthquake damage assessment of buildings using VHR optical and SAR imagery IEEE T GEOSCI REMOTE vol 48 pp 2403-2420
[5] Ramirez C M, Liel A B, Mitrami-Reiser J, Krawinkler H, Haselton C B, Spear A, Steiner J, Deierlein G G and Miranda E 2012 Expected earthquake damage and repair costs in reinforced concrete frame buildings EARTHQ ENG STRUC TD vol 41 pp 1455-1475
[6] Zhai W, Shen H, Huang C and Pei W 2016 Building earthquake damage information extraction from a single post-earthquake PolSAR image REMOTE SENS-BASEL vol 8 p 171
[7] Wu J, He X, Ye M and Wang C 2019 Energy and asset value elasticity of earthquake-induced direct economic losses INT J DISAST RISK RE vol 33 pp 229-234
[8] Rashid M and Ahmad N 2017 Economic losses due to earthquake-induced structural damages in RC SMRF structures COGENT ENG vol 4 pp 1
[9] Sadeghi M, Hochrainer-Stigler S and Ghafoory-Ashtiany M 2015 Evaluation of earthquake mitigation measures to reduce economic and human losses: a case study to residential property owners in the metropolitan area of Shiraz, Iran NAT HAZARDS vol 78 pp 1811-1826
[10] Lallemant D, Kiremidjian A and Burton H 2015 Statistical procedures for developing earthquake damage fragility curves EARTHQ ENG STRUCTD vol 44 pp 1373-1389
[11] Anniballe R, Noto F, Scalia T, Bignami C, Stramondo S, Chini M and Pierdicca N 2018 Earthquake damage mapping: an overall assessment of ground surveys and VHR image change detection after L’Aquila 2009 earthquake REMOTE SENS ENVIRON vol 210 pp 166-178
[12] Liu R, Yu S, Yan D, He M and Yue X 2014 Refined Method for Rapid Assessment of Seismic Damage and Economic Loss J BASIC SCI ENG vol 22 pp 928-940 (In Chinese)
[13] Finch K C, Snook K R, Duke C H, Fu K, Tse Z, Adhikari A and Fung I C 2016 Public health implications of social media use during natural disasters, environmental disasters, and other environmental concerns NAT HAZARDS vol 83 pp 729-760
[14] Lachlan K A, Spence P R, Lin X, Najarian K M and Greco M D 2016 Social media and crisis management: CERC, search strategies, and Twitter content COMPUT HUM BEHAV vol 54 pp 647-652
[15] Avvenuti M, Cresci S, Marchetti A, Meletti C and Tesconi M 2016 Predictability or early warning: using social media in modern emergency response IEEE INTERNET COMPUT vol 20 pp 4-6
[16] Cresci S, Avvenuti M, La Polla M, Meletti C and Tesconi M 2017 Nowcasting of earthquake consequences using big social data IEEE INTERNET COMPUT vol 21 pp 37-45
[17] Shu K, Sliva A, Wang S, Tang J and Liu H 2017 Fake news detection on social media: A data mining perspective ACM SIGKDD EXPLOR NEWSL vol 19 pp 22-36
[18] Qian J, Li X Y, Zhang C, Jung T and Chen L 2017 Social network de-anonymization and privacy inference with knowledge graph model IEEE T DEPEND SECURE vol 16 pp 679-692
[19] Gottschalk S and Demidova E 2018 Eventkg: A multilingual event-centric temporal knowledge graph European Semantic Web Conf. (Greece) (Lecture Notes in Computer Science vol 10843) ed Gangemi A (Cham: Springer) pp 272-287
[20] Li J and Hou L 2017 Reviews on Knowledge Graph Research J SHANXI UNIV (Nat. Sci. Ed.) vol 40 pp 454-459 (In Chinese)
[21] Yuan X 2015 Construction of Business Knowledge Graph Based on Deep Learning (Shanghai: East China Normal University) (In Chinese)
[22] Proux D, Rechenmann F, Julliard L, Pillet V and Jacq B 1998 Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction GENOME INFORM vol 9 pp 72-80
[23] Rau L F 1991 Extracting company names from text Proc. The 17th IEEE Conf. on Artificial Intelligence Application ed Poh H L (Piscataway: IEEE) pp 29-32
[24] Liu Q, Li Y, Duan H, Liu Y and Qin Z 2016 Knowledge graph construction techniques J COMPUT RES DEVELOP vol 53 pp 582-600 (In Chinese)
[25] Cambria E and White B 2014 Jumping NLP Curves: A Review of Natural Language Processing Research IEEE COMPUT INTELL M vol 9 pp 48-57
[26] Xu Z and Li Y 2017 Create a library of NLP text classification deep learning methods ROBOT IND vol 3 pp 50-57 (In Chinese)
[27] Lample G, Ballesteros M, Subramanian S, Kawakami K and Dyer C 2016 Neural Architectures for Named Entity Recognition Preprint 1603.01360
[28] Luo L, Yang Z, Yang P, Zhang Y, Wang L, Lin H and Wang J 2017 An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition Bioinformatics vol 34 pp 1381-1388 (In Chinese)
[29] Qin Y, Shen G, Zhao W, Chen Y, Yu M and Jin X 2019 A network security entity recognition method based on feature template and CNN-BiLSTM-CRF FRONT INFORM TECH EL vol 20 pp 872-884 (In Chinese)
[30] http: //colah. github. io/posts/2015-08-Understanding-LSTMs
[31] Bale T L and Vale W W 2004 CRF AND CRF RECEPTORS: Role in Stress Responsivity and Other Behaviors ANNU REV PHARMACOL vol 44 p 525
[32] Lu H, Hong Z and Shi M 2017 Analysis of film data based on Neo4j IEEE/ACIS Int. Conf. on Computer & Information Science (Wuhan) ed Shang W (Piscataway: IEEE) pp 675-677
[33] https://www.cnblogs.com/Determined22/p/6915730.html
[34] Lin Y, Liu Z, Sun M, Yiu L and Zhu X 2015 Learning entity and relation embeddings for knowledge graph completion Proc. of the 29th AAAI Conf. on Artificial Intelligence (Texas) ed Cohn A (Palo Alto: AAAI Press) vol 3 pp 2181-218