A Practice of Tourism Knowledge Graph Construction based on Heterogeneous Information

Dinghe Xiao
Hainan Sino-intelligent-Info Technology Ltd.
xiaodinghe@e-zzx.com

Nannan Wang
Beijing University of Posts and Telecommunications
wangnannan@bupt.edu.cn

Jiangang Yu
Hainan Sino-intelligent-Info Technology Ltd.
yujiangang@e-zzx.com

Chunhong Zhang✉
Beijing University of Posts and Telecommunications
zhangch@bupt.edu.cn

Jiaqi Wu
Hainan Sino-intelligent-Info Technology Ltd.
wujiaqi@e-zzx.com

Abstract

The increasing amount of semi-structured and unstructured data on tourism websites brings a need for information extraction (IE) so as to construct a Tourism-domain Knowledge Graph (TKG), which is helpful to manage tourism information and develop downstream applications such as tourism search engine, recommendation and Q & A. However, the existing TKG is deficient, and there are few open methods to promote the construction and widespread application of TKG. In this paper, we present a systematic framework to build a TKG for Hainan, collecting data from popular tourism websites and structuring it into triples. The data is multi-source and heterogeneous, which raises a great challenge for processing it. So we develop two pipelines of processing methods for semi-structured data and unstructured data respectively. We refer to tourism InfoBox for semi-structured knowledge extraction and leverage deep learning algorithms to extract entities and relations from unstructured travel notes, which are colloquial and high-noise, and then we fuse the extracted knowledge from two sources. Finally, a TKG with 13 entity types and 46 relation types is established, which totally contains 34,079 entities and 441,371 triples. The systematic procedure proposed by this paper can construct a TKG from tourism websites, which can further applied to many scenarios and provide detailed reference for the construction of other domain-specific knowledge graphs.

1 Introduction

Tourism has become increasingly popular in people's daily life. Before people set out to travel, they often need to make clear the travel guides and matters needing attention for their destinations. Nowadays, with the development of the Internet, many tourism websites have appeared and provide a variety of travel information, such as attractions, tickets, bus routes, travel guides, etc. However, there may be some errors in the miscellaneous information on the tourism websites, and information on different tourism websites may be inconsistent. As shown in screenshots of Sina Micro-Blog users' blogs in Figure 1, there are still tourists who are worried about making travel strategies despite rich information on all kinds of tourism-related search engines. How to collect and integrate valuable tourism knowledge on websites is a very important issue.

Recently, Knowledge Graph (KG) has received much attention and research interest in industry and academia. The KG utilizes a set of subject-predicate-object triplets to represent the diverse entities and their relations in real-world scenes, which are respectively represented as nodes and edges in the graph. The KG is a graph-based large-scale knowledge representation and integration method, which has been applied in various scenarios such as enterprise (Miao et al., 2015), medical (Rotmensch et al., 2017) and industry (Zhao et al., 2019). Naturally, we consider applying KG in the field of Tourism to integrate and organize relevant knowledge, so as to provide tourists with easier tools to develop travel strategies.

At present, several General Knowledge Graphs (GKGs) have been built both in Chinese and English (Auer et al., 2007; Suchanek et al., 2007; Niu et al., 2011; Xu et al., 2017). The Domain-specific

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Figure 1: Screenshots of Sina Micro-Blog users' blogs. In the blogs, people with tourism intentions complain that it is difficult to formulate travel strategies.

Knowledge Graph (DKG) in which the stored knowledge is limited to a certain field has also been implemented and put into use in many domains (Zhao et al., 2018). However, Tourism-domain Knowledge Graph (TKG) is still deficient, which undoubtedly hinders the development of intelligent tourism system. In this paper, we propose a systematic framework to construct a TKG under the background of Hainan Tourism. We combine the semi-structured knowledge crawled from the encyclopedia pages of tourism websites with the unstructured travel notes shared by tourists on the websites as the data source. Because of the lack of sufficient high-quality data and the difficulty of language processing, constructing a Chinese-based TKG still faces several challenges as follows:

Travel notes are colloquial and high-noise. The writing style of travel notes is often arbitrary, and tourists tend to add various pictures, emoticons and special characters to travel notes, which will introduce much noise for unstructured data.

The Lack of datasets dedicated to tourism. There is a serious lack of normative datasets in the tourism field, which are basis of model training.

Are the general algorithms suitable for tourism? Entity extraction and relation extraction are the key steps in knowledge graph construction. Most of the existing algorithms for these two tasks are tested on the general datasets, we need to verify whether these algorithms are suitable for the tourism field.

How to integrate data from different sources? Data from different sources inevitably have some overlaps and ambiguities, which should be eliminated in the KG.

Facing this challenges, we put forward corresponding methods to deal with them. In detail, the contributions of our work are highlighted as follows:

- A specific method of collecting and processing tourism-domain data is described, and labeled datasets for information extraction in the field of tourism is constructed;
- The most suitable models for our tourism data are identified, and a tourism-domain knowledge graph is finally constructed.
- Experience in constructing the TKG can provide detailed reference for the construction of other domain-specific knowledge graphs.

2 Related Work

In recent years, the KG has been applied in many fields to complete knowledge storage, query, recommendation and other functions. In the tourism scene, experts and scholars have also begun to explore the application value of knowledge graphs. DBtravel (Calleja et al., 2018) is an English tourism-oriented knowledge graph generated from the collaborative travel site Wikitravel. A Chinese TKG was also constructed by (Zhang et al., 2019), which extracted tourism-related knowledge from existing Chinese

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

general knowledge graph such as zhishi.me (Niu et al., 2011) and CN-DBpedia (Xu et al., 2017). Unlike their Chinese TKG, we extensively obtain data and extract knowledge from popular tourism websites. In this way, the completeness of our knowledge graph does not depend on the existing knowledge graph, but on the amount of data we acquire. To construct the TKG, we need to extract triples form all kinds of information resources. The conversion process from semi-structured data to structured data is more standardized and has fewer errors, but semi-structured data often cannot contain all the knowledge. With the development of Natural Language Processing (NLP), more and more knowledge graphs are constructed based on unstructured corpus, using named entity recognition (NER) and relation extraction (RE) technologies.

As a hot research direction in the field of NLP, many Chinese NER models have been proposed over the years. The purpose of NER task is to identify mentions of named entities from text and match them to pre-defined categories. As a classic branch of NER models, the dictionary-based methods recognize named entities by constructing a dictionary and matching text with it. For example, CMEL (Meng et al., 2014) built a synonym dictionary for Chinese entities from Microblog and adopts improved SVM to get textual similarity for entity disambiguation. Another line of related work is to apply traditional machine learning techniques to complete the NER task, just like the Conditional Random Fields(CRFs)-based NER System proposed by (Han et al., 2013). Recently, neural network-based (NN-based) models have shown great future prospects in improving the performance of NER systems, including bidirectional Long Short-Term Memory (LSTM) model (He et al., 2019), lattice-structured LSTM model (Zhang and Yang, 2018), convolution neural network (CNN)-based model (Gui et al., 2019) and so on. In our work, we adopt the most mainstream NN-based NER algorithm at present, which combines BiLSTM and CRF.

Relation extraction (RE) is also one of the most important tasks in NLP. On the premise of pre-defined relation categories, RE is often transformed into a relation classification task. Similar to entity extraction, the mainstream algorithms for RE in recent years have also focused on NN-based ones. Zeng et al. (2014) utilized CNNs to classify relations and made representative progress. However, because CNN can not extract contextual semantic information well, recurrent neural network (RNN) (Zhang and Wang, 2015), which is often used to process texts, is proposed for relation extraction. Since RNN is difficult to learn long-term dependencies, LSTM (Zhang et al., 2015) was introduced into the RE task. To capture the most important information in a sentence, Attention-Based Bidirectional Long Short-Term Memory Networks(Att-BLSTM) (Zhou et al., 2016) was come up and become a popular RE algorithm. The above supervise learning algorithms are time-consuming and costly to label data. In order to solve these problems, some distant supervision algorithms have also been developed (Zeng et al., 2015; Han and Sun, 2016; Ji et al., 2017). Because the TKG only contains knowledge in the field of tourism, the corpus for training is not large, so we do not consider using distant supervision algorithms.

3 Implementation

In this paper, we crawl semi-structured and unstructured data related to Hainan Tourism from popular travel websites, and extract the structured knowledge from these two types of data in two pipelines. Figure 2 shows the overview of our method.

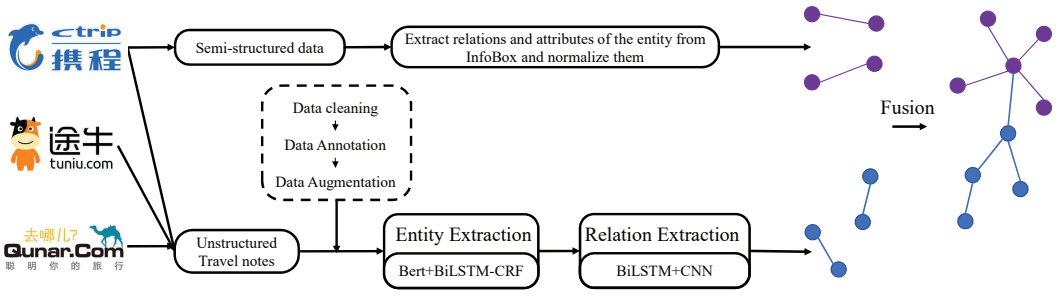


Figure 2: The overview of our method.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.1 Data Preparation

Tourism is an intelligent application market with great potential. Tourism data on the Internet has a large quantity but not effectively used, and standardized tourism datasets are not yet available. In this section, we will describe our data preparation process in detail, which is mainly divided into four steps including data acquisition, data cleaning, data annotation and data augmentation, and the last three steps are mainly aimed at unstructured data that is noisy and irregular.

Data acquisition: This step aims to collect raw data in the field of tourism, which will be processed later to be used as input to the information extraction models. There are many popular Chinese tourism websites that cover numerous tourism-related knowledge on the Internet. We crawled semi-structured data on the Ctrip[0], where tourism-related entities (scenic areas, hotels, cities, etc.) have their corresponding descriptive pages. The Information Boxes (InfoBox) in these pages with clear structure contain a great number of named entities, relations and attributes, which can be used to fill the TKG. For example, the InfoBox of "Haikou Ublaya Inn" is shown in the Figure 3(a). Meanwhile, we crawled tourists' travel notes related to Hainan on the three major Chinese travel websites, Ctrip[1], Tuniu[2] and Qunar[3]. Travel notes are rich in content and easy to obtain, which may supplement the information not contained in semi-structured data, and Figure 3(b) shows an example of travel notes on the Tuniu.

(a) InfoBox Example. (b) Travel notes Example.

Figure 3: An example of (a) an InfoBox of "Haikou Ublaya Inn" and (b) travel notes related to Hainan on the Tuniu, which respectively correspond to the semi-structured data and unstructured data that we want to crawl on the travel websites.

We have crawled 33177 pages corresponding to Hainan-related entities from the Ctrip. In addition, a total of 19,023 travel notes are obtained after crawling the above three popular websites. The combination of semi-structured data and unstructured data helps to provide a more complete source of information in the construction of TKG.

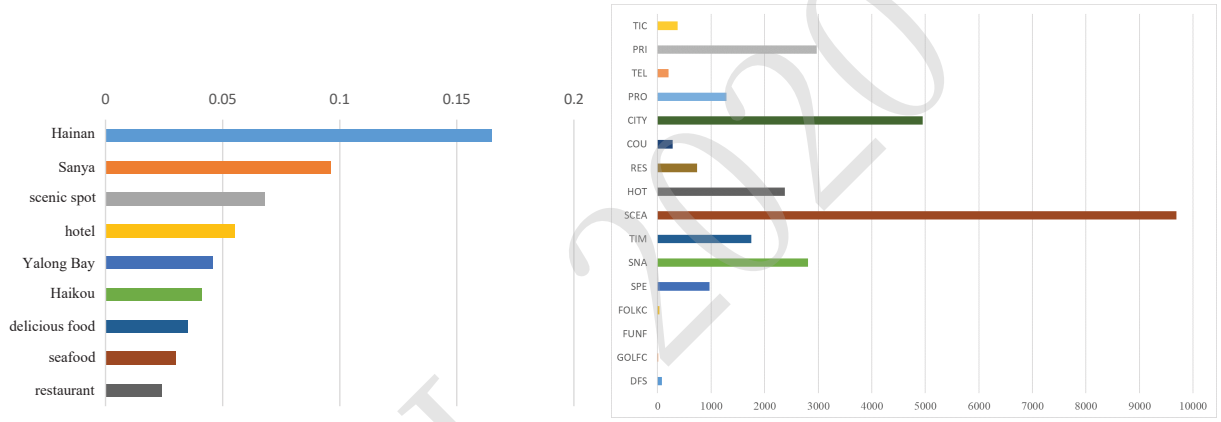
Data cleaning: For unstructured data, due to the colloquial and casual nature, the travel notes crawled from the travel websites usually contain some noise that should be cleaned up, including some inconsistent Traditional Chinese characters, emoticons, Uniform Resource Locator (URL) links and some special characters like #, &, $, {, }, etc. We mainly delete these redundant contents through regular expressions. In view of the fact that some paragraphs in travel notes are relatively longer than the ideal length required by the models for entity extraction and relation extraction, we further perform paragraph segmentation to reduce the pressure of model training.

Data Annotation: For unstructured text, we should label it to build datasets that meet the training requirements for subsequent entity recognition and relation recognition algorithms. Before annotating data, we must first define the types of entities and relations that need to be extracted in the field of tourism. In order to truly understand the issues that users are concerned about, we crawl the text about

[0]https://you.ctrip.com/place/100001.html
[1]https://you.ctrip.com/travels/
[2]https://trips.tuniu.com/
[3]https://travel.qunar.com/

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the keyword ”Hainan” in the QA modules of Ctrip and Tuniu, mainly including some users' questions and the answers given by other users, and then the word frequency in the Q & A data is analyzed through TF-IDF (Term Frequency-Inverse Document Frequency) algorithm. The statistical results of word frequency in our work are shown in figure 4(a). The results show that high-frequency words are mainly concentrated on types such as hotel, scenic spot, city, food, restaurant, etc. Referring to the definition of entities and relations in CN-DBpedia (Xu et al., 2017), we define 16 entity types and 51 relation types that should be extracted from unstructured data based on the features of tourism-domain data. Entity types include DFS (Duty Free Shop), GOLFC (Golf Course), FUNF (Funfair), HOT (Hotel), FOLKC (Folk Custom), SPE (Specialty), SNA (Snacks), TIM (Time), TEL (Telephone), PRI (Price), TIC (Ticket), SCEA (Scenic Area), PRO (Province), CITY (City), COU (County) and RES (Restaurant). Because of the relatively large number of relation types, we give an example to illustrate the relation types. When choosing a restaurant, tourists need to figure out the location, price, business hours and phone number of the hotel, and the location must be specific. So we define seven relations for RES type, including res_locatedin_scea, res_locatedin_pro, res_locatedin_city, res_locatedin_cou, res_open_time, res_phonenumber, res_PRI, where res_locatedin_scea means that the restaurant is in a certain scenic area, and the explanation of the remaining relations is similar.



(a) Statistics of word frequency.

(b) Statistics of entities of 16 types.

Figure 4: (a) Word frequency statistics in the Q & A data, where high-frequency words need to be focused on; (b) Statistics of the number of 16 types of entities, it shows that the number of entities is unevenly distributed.

After defining the entity & relation types to be extracted, for a sentence in travel notes, we should first label entity mentions in it, and then label the relation between entity pairs according to semantics. We adopt BRAT (Stenetorp et al., 2012) as the main tool to label entities and relations in the text. There exist some problems when using BRAT to label entities and relations in the field of tourism. When labeling entities, 1) The travel notes are expressed by different people in a colloquial way, which makes it difficult to determine the boundary of the entities. We reasonably label the entity mentions with the boundary as large as possible, so as to make the entity mention more complete and specific; 2) In different contexts, entities with the same mention may belong to different types. So we label relations based on the semantics of the context. There are also some problems when labeling relations, 1) When multiple entities appear in a sentence, and there is more than one entity pair that has connections, we label as many entity-relation-entity combinations to obtain adequate relation annotated data; 2)A sentence may contain two entities, and there may be a connection between the two entities according to external knowledge, but the context of the sentence cannot reflect this connection. For this situation, we will not label the relation, so as not to have a negative impact on the subsequent training of the RE model.

After handling the above problems, 1902 travel notes are annotated. Because labeling relations needs to consider the context, which affects the speed of the labeling, we have not labeled all crawled travel notes, but only labeled the number enough to train the models. The details of the datasets will be shown

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

in Section 4.1.

Data Augmentation: The number of entities in travel notes is not evenly distributed in categories. We make statistics on the number of entities of each entity type contained in the annotated dataset, as shown in the figure 4(b). We can see that there are a large number of labeled entities in SCEA and CITY types, and the proportion of other types is relatively small. In order to reduce the training error brought by data imbalance, we use substitution method to expand the types with a small amount of data. We take the DFC entities with a small proportion as example. First, select some sentences containing the DFC entity from the dataset, and then replace the DFC entity mentions in each sentence with other different DFC mentions. Although such replacement destroys the authenticity of the original data, the training for models is appropriate. We use this technology to augment a total of more than 8,000 pieces of sentence.

3.2 Knowledge Extraction of Semi-structured Data

Since a page crawled on Ctrip tends to contain the description of the relevant attributes and relations of only one named entity, we extract not only entity mention but also the corresponding URL, and the URL can uniquely represent the entity. In this way, we successfully extract the uniquely identifiable entities from the semi-structured data, and there is no ambiguity between these entities. In addition, we extract attributes and relations of a entity mainly through the InfoBox. It is worth noting that there are many cases of inconsistent attributes and value conflicts in Semi-structured data. For example, attribute names can be inconsistent (telephone, contact number), and attribute values can be inconsistent(086-6888-8888 and 68888888), so for the extracted semi-structured knowledge, we further refer to CN-DBpedia (Xu et al., 2017) for attribute normalization and value normalization to further obtain well-organized knowledge, and then we finally obtain about 370,000 triples from semi-structured data.

3.3 Knowledge Extraction of Unstructured Data

In this section, we mainly utilize mainstream deep learning algorithms to extract entities and relations in unstructured text.

Entity extraction. For unstructured data from tourist travel notes, we take the method of Named Entity Recognition (NER) to extract entity mentions from the text. The main work of NER is sequence labeling, and Long Short-Term Memory (LSTM) networks have natural advantages in processing time series related tasks. The Conditional Random Field (CRF) model can effectively consider the mutual influence of output labels between characters. Therefore, the BiLSTM model and the CRF model are usually used together to become the mainstream model in the NER field.

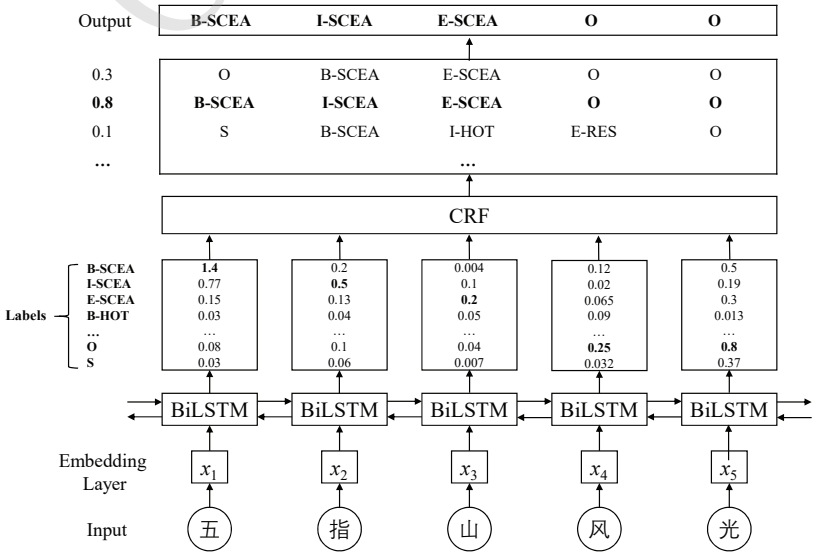


Figure 5: The baseline framework of entity extraction model based on BiLSTM-CRF.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The BiILSTM-CRF (Huang et al., 2015) model diagram is shown in Figure 5. The input in English is "Five-Finger Mountains' scenery", and the output means that Five-Finger Mountains belong to the entity type SCEA. Next, We make further improvements in the embedding layer. After the Google BERT (Devlin et al., 2018) model was proposed, the innovation of the pretrained language model has enabled many NLP tasks to achieve state-of-the-art performance, and large pretrained language models have become a hot tool. After BERT, other large pretrained language models like ALBERT (Lan et al., 2019) model have also been proposed. ALBERT is a simplified BERT version, and the number of parameters is much smaller than the traditional BERT architecture. In this paper, we utilize the pretrained BERT and ALBERT model to obtain the embedding matrix in embedding layer respectively, which is constant during the training process.

Relation extraction. Relation Extraction (RE) is an important task of natural language processing (NLP) and also a key link in knowledge graph construction. After RE, a triple (s, r, o) is usually obtained, where s represents the head entity, o represents the tail entity, and r represents the relation between them. In our travel data, the number of relations is limited, so we can choose to transform the RE into a relation classification task, and we treat each relation type as a class. Comprehensively considering the advantages and disadvantages of the mainstream relationship classification model and characteristics of tourism data, we choose to adopt supervised algorithm, BiLSTM+CNN (Zhang and Xiang, 2018), for RE task in our work, whose framework can be shown in Figure 6. CNN can extract local features of sentences, but it is not good at handling long dependencies among words, which can be made up by BiLSTM.

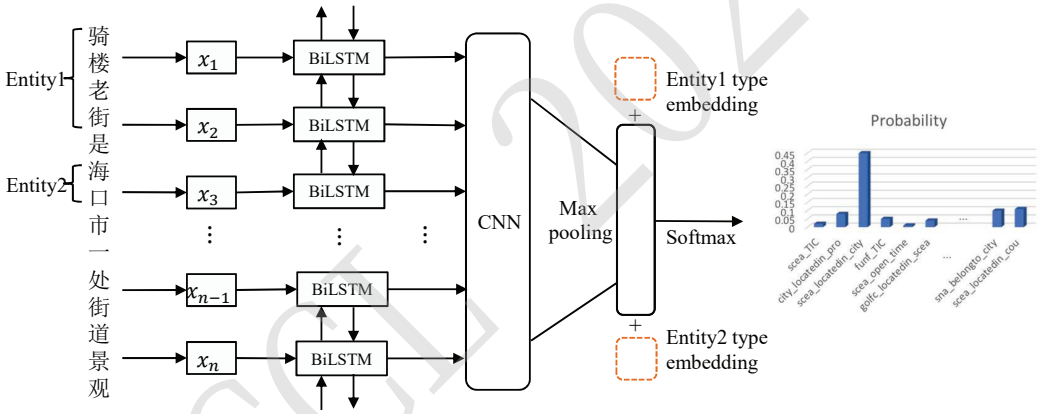


Figure 6: The framework of the relation extraction model based on BiLSTM+CNN. The input in English is "Qilou Old Street is a street view of Haikou City.", and from the bar graph we know that the output relation is scea_locatedin_city, then we can get the triple (Qilou Old Street, scea_locatedin_city, Haikou).

At the same time, considering that the entity category information may have an impact on the relation classification, the entity type information is introduced into the model (Lee et al., 2019). Specifically, each entity type is represented as distributed embedding. As is shown in Figure 6, after the CNN layer, we concatenate the entity type embedding of entity1 and entity2 with the output vector of Max pooling layer, and then feed it to the fully connected layer for subsequent label prediction.

Entity Alignment. There is often a situation where multiple mentions refer to the same entity. Entity alignment is to determine whether two entities with different mentions are the same entity by calculating and comparing their similarity. We observe the names of the entities that need to be aligned and find that the names of the two entities to be aligned are similar in most cases, just like "Nantian Ecological Grand View Garden" and "Nantian Grand View Garden". Therefore, basic distance measurement-based models are suitable enough for our entity alignment task, which is to calculate the distance between the names of the two entities. Common distance measurement algorithms include Jaccard coefficient, Euclidean distance, and editing distance. We weight and sum the distances measured under these three distance metrics, so as to discriminate whether entities with different names belong to the same entity. Although this method is simple, but it can solve most of the problems we encounter. Finally, we obtain about

220,000 triples from unstructured data.

In summary, we first construct independent knowledge graphs from two heterogeneous data sources respectively, and then we fuse the two sub-knowledge graphs to obtain a more complete knowledge graph, which is the Tourism-domain Knowledge Graph finally constructed.

4 Experiments

4.1 Datasets

In Section 3.1, we acquire, clean, annotate and augment the unstructured text crawled from popular travel websites, and obtain two labeled datasets suitable for Named Entity Recognition (NER) and Relation Extraction (RE) tasks. For labeled datasets, post-processing operations are needed to eliminate data that is meaningless for model training. Specifically, if there is no entity in a sentence, delete it directly. If the sentence contains only one entity, it will be cut to the proper length and only be used for NER training. Our datasets are both based on sentences, and a sentence is a piece of data. For NER dataset, we use train, valid, and test splits of 5490, 1178, and 591 sequence labeled sentences respectively. And train, valid, and test sets for RE task contain 6225, 1000 and 400 sentences respectively. Using the datasets we construct and divide, we next conduct comparative experiments to measure the model performance.

4.2 Model Training and Results

In order to obtain a named entity recognition model suitable for tourism-domain data, we compare several mainstream NER models including BERT(Cai, 2019), ALBERT(Lan et al., 2019), BiLSTM-CRF(Huang et al., 2015), BERT+BiLSTM-CRF(Dai et al., 2019), and BERT-CRF(Souza et al., 2019) on our NER dataset. For this task, we use Precision (P), Recall (R) and F1 score (F1) to evaluate the effect of NER model, which are standard information extraction metrics. The experimental results in Table 1 show that the BiLSTM-CRF algorithm based on the pretrained language model BERT has the best performance with F1-score 90.6%. BERT+BiLSTM-CRF practiced by Dai et al. (2019) is used to complete the task of Chinese electronic medical records named entity recognition, and BERT+BiLSTM-CRF achieves approximately 75% F1 score and performs better than other models like BiLSTM-CRF and BiGRU-CRF in their work, which is consistent with our results. Both in their and our practice, the effectiveness of combining pretrained models with mainstream models is reflected. Meanwhile, we can see that baselines other than BERT+BiLSTM-CRF that have good performance on the general standard datasets can also achieve comparative results in the application of actual projects.

The NER models share the same divided NER dataset and training environment, and all models are trained with 15 epochs.

	Model	P	R	F1
NER	BiLSTM-CRF(Huang et al., 2015)	**0.890**	0.876	0.883
	BERT-CRF(Souza et al., 2019)	0.862	0.904	0.882
	BERT(Cai, 2019)	0.822	0.867	0.839
	ALBERT(Lan et al., 2019)	0.837	0.829	0.828
	BERT+BiLSTM-CRF(Dai et al., 2019)	0.887	**0.926**	**0.906**
RE	BiLSTM+ATT(Zhou et al., 2016)	0.766	0.681	0.702
	CNN(Zeng et al., 2014)	0.803	0.651	0.701
	BiLSTM-CNN(Zhang and Xiang, 2018)	**0.941**	0.791	0.842
	BiLSTM-CNN(with types)	0.918	**0.914**	**0.909**

Table 1: Comparison of experimental results with NER baselines and RE baselines on our datasets.

Similar to entity extraction, we also compare three mainstream models in relation extraction task, including BiLSTM+ATT (Zhou et al., 2016), CNN (Zeng et al., 2014) and BiLSTM-CNN (Zhang and Xiang, 2018). The evaluation metrics applied in RE models are also P, R and F1. Among these models, as is shown in Table 1, BiLSTM-CNN shows the relatively better performance than BiLSTM+ATT and CNN on our RE dataset.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

In order to further verify the validity of adding entity type embedding in RE, comparative experiments are carried out on the model BiLSTM-CNN. Table 1 shows that by introducing entity type information, the F1 score of BiLSTM-CNN is improved by 6.7%, which is the highest among our experimental models. The main reason may be that by introducing entity type information into the model, the scope of classification is narrowed, that is to say, entity type information restricts the classification to a certain extent, so as to significantly improve the effect of relation classification. The above RE models share the same divided RE dataset and training environment, and all models are trained with 64 epochs.

To sum up, based on the above analysis of the experimental results of each model, BERT+BiLSTM-CRF is selected as NER model and BiLSTM + CNN model with entity type information introduced is selected as the RE model in our work.

4.3 Knowledge Construction

We fuse the two sub-knowledge graphs obtained from semi-structured data and unstructured data to get the complete TKG. The final TKG with a total of 441,371 triples contains 13 entity types and 46 relation types. In Figure 7, a knowledge graph composed of partial triples is depicted. The central node is Sansha that belongs to CITY type, and we show a part of the nodes around it and the adjacent relations and attributes.

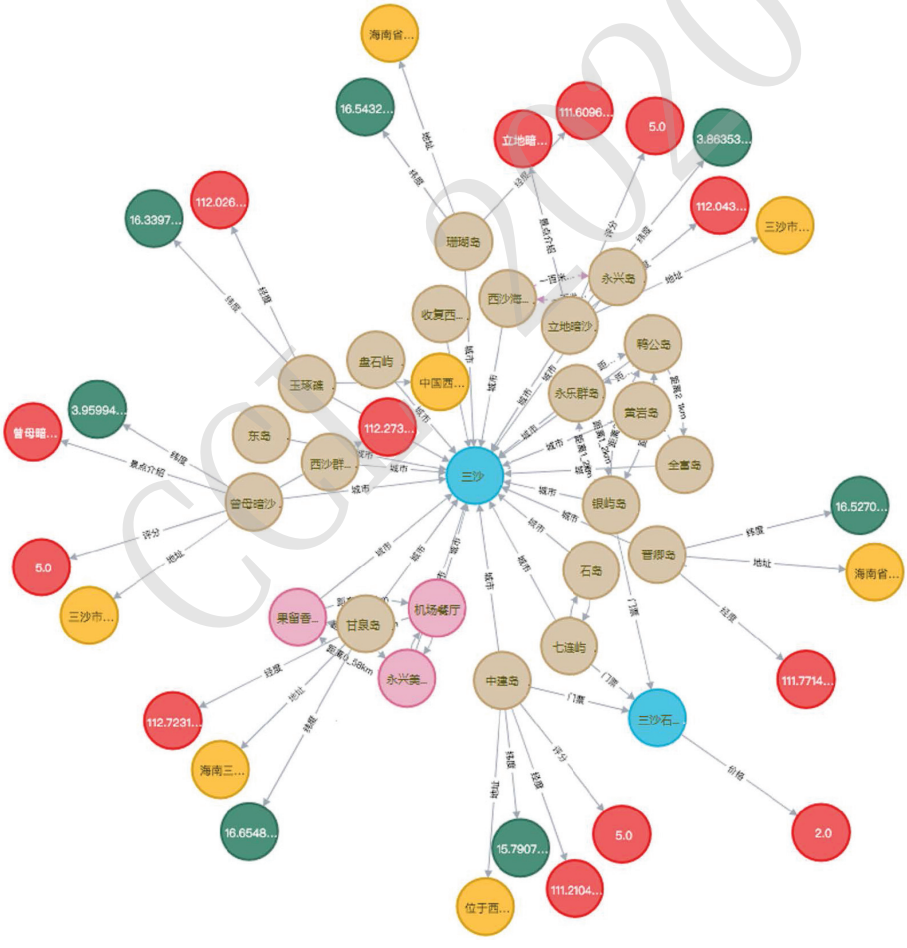


Figure 7: Partial triples in tourism knowledge graph, which shows the part of the tourism-domain knowledge graph with CITY Sansha as the central node.

5 CONCLUSIONS

With the development of tourism, information management and utilization in the field of tourism is a very important task. We proposed a systematic approach to construct the Chinese tourism knowledge graph,

using the information on the tourism websites. We leveraged semi-structured data and unstructured data to extract entities and relations synchronously, and they can be combined to obtain more complete sets of entities and relations than only one of them. Due to the lack of standardized datasets in the field of tourism, we first proposed a strategy for constructing datasets to facilitate the extraction of entities and relations from the complex network text data. In addition, we used several algorithms to complete the named entity recognition (NER) task and relation extraction (RE) task on the datasets we created, and compare the results. We found that BERT+BILSTM-CRF has the best performance for NER task and BiLSTM+CNN with entity type information introduced performs best on RE task.

We have implemented a relatively complete information extraction system on the tourism knowledge graph. In the future work, we want to solve the problem of how to update the knowledge in real time, because the knowledge on the tourism websites is always increasing and changing. In addition, we intend to explore some domain-adaptive techniques to make our model can be used widely.

Acknowledgements

This work is supported by the Science and Technology Department of Hainan Province, "Intelligent analysis platform of Hainan tourist's behavior and accurate service mining prediction" project(ZDKJ201808).

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Qing Cai. 2019. Research on chinese naming recognition model based on bert embedding. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pages 1–4. IEEE.

Pablo Calleja, Freddy Priyatna, Nandana Mihindukulasooriya, and Mariano Rico. 2018. Dbtravel: A tourism-oriented semantic graph. In *International Conference on Web Engineering*, pages 206–212. Springer.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4982–4988. AAAI Press.

Xianpei Han and Le Sun. 2016. Global distant supervision for relation extraction. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Aaron L-F Han, Derek F Wong, and Lidia S Chao. 2013. Chinese named entity recognition with conditional random fields in the light of chinese characteristics. In *Intelligent Information Systems Symposium*, pages 57–68. Springer.

Zhonghe He, Zhongcheng Zhou, Liang Gan, Jiuming Huang, and Yan Zeng. 2019. Chinese entity attributes extraction based on bidirectional lstm networks. *International Journal of Computational Science and Engineering*, 18(1):65–71.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Joohong Lee, Sangwoo Seo, and Yong Suk Choi. 2019. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry*, 11(6):785.

Zeyu Meng, Dong Yu, and Endong Xun. 2014. Chinese microblog entity linking system combining wikipedia and search engine retrieval results. In *Natural Language Processing and Chinese Computing*, pages 449–456. Springer.

Qingliang Miao, Yao Meng, and Bo Zhang. 2015. Chinese enterprise knowledge graph construction based on linked data. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 153–154. IEEE.

Xing Niu, Xinruo Sun, Haofen Wang, Shu Rong, Guilin Qi, and Yong Yu. 2011. Zhishi. me-weaving chinese linking open data. In *International Semantic Web Conference*, pages 205–220. Springer.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. 2017. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):1–11.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 428–438. Springer.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Lei Zhang and Fusheng Xiang. 2018. Relation classification via bilstm-cnn. In *International Conference on Data Mining and Big Data*, pages 373–382. Springer.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 73–78.

Weizhen Zhang, Han Cao, Fei Hao, Lu Yang, Muhib Ahmad, and Yifei Li. 2019. The chinese knowledge graph on domain-tourism. In *Advanced Multimedia and Ubiquitous Engineering*, pages 20–27. Springer.

Zhanfang Zhao, Sung-Kook Han, and In-Mi So. 2018. Architecture of knowledge graph construction techniques. *International Journal of Pure and Applied Mathematics*, 118(19):1869–1883.

Mingxiong Zhao, Han Wang, Jin Guo, Di Liu, Cheng Xie, Qing Liu, and Zhibo Cheng. 2019. Construction of an industrial knowledge graph for unstructured chinese text learning. *Applied Sciences*, 9(13):2720.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 939-949,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China