Natural Language Processing and Speech Recognition (2019) 1: 1-5
Clausius Scientific Press, Canada

DOI: 10.23977/nlpsr.2019.11001
ISSN 2523-9082

A Survey of Research on Deep Learning Entity Relationship Extraction

Wang Peng

School of Economics and Management, Dalian University, No.10, Xuefu Avenue, Economic & Technical Development Zone, Dalian, Liaoning, The People's Republic of China(PRC)

Keywords: Entity relationship extraction, Deep learning, Joint learning, Remote supervision, Generating confrontation network

Abstract: Entity relationship extraction is the core task and important link in the fields of information extraction, natural language understanding, information retrieval, etc. It can extract the semantic relationship between entity pairs from text. In recent years, the application of deep learning in joint learning and remote supervision The relationship extraction task has obtained rich research results. At present, the deep learning-based entity relationship extraction technology has gradually surpassed the traditional feature-based and kernel-based methods in feature extraction depth and model accuracy. And the two areas of remote supervision, the system summarizes the research progress of Chinese and foreign scholars' deep relationship-based entity relationship extraction in recent years, and discusses and prospects the future research directions.

1. Introduction

With the development of Internet technology, the amount of data that people need to process has proliferated, and the phenomenon of cross-domain has become prominent. How to extract effective information from open-field text quickly and efficiently becomes an important issue in front of people. Entity relationship extraction as text mining And the core task of information extraction,Mainly by modeling the text information, the semantic relationship between the entity pairs is automatically extracted, and the effective semantic knowledge is extracted. The research results are mainly applied to text abstract, automatic question and answer, machine translation, semantic web annotation, knowledge map [4]. Etc. With the rise of information extraction in recent years, the issue of entity relationship extraction has received further attention and in-depth research. Some research results have appeared in international conferences in related fields such as artificial intelligence and natural language processing in recent years, ACL, EMNLP. , ICLR, AAA, KDD, NAACL, ECML-PKDD, etc.

The classical method has the problem of feature extraction error propagation, which greatly affects the effect of entity relationship extraction. With the rise of deep learning in recent years, scholars have gradually applied deep learning to the task of entity relationship extraction. Based on the difference of data set dimension, Deep learning entity relationship extraction tasks are divided into There are two types of supervised and remote supervision. The supervised entity relationship extraction method based on deep learning is a research hotspot of relationship extraction in recent years. This method can avoid the steps of artificial feature selection in the classical method, and reduce and improve the error accumulation in the feature extraction process. Question. According to the order of completion of the two subtasks of entity identification and relationship classification, the method of extracting supervised entity relations based on deep learning can be divided into pipeline method and joint learning method. Zeng first proposed in 2014. Use CNN for relationship classification, Katiyar in 2017 For the first time, attention attention mechanism Attention and recurrent neural network Bi-LSTM are used to jointly extract entity and classification relations. The neural network model has achieved good results in the development of supervised fields. At the same time, the method of remote monitoring entity relationship extraction based on deep learning It has become a research hotspot with the ability to alleviate the problem of error labeling and feature extraction error propagation in classical methods. The main basic methods include CNN, RNN, LSTM and other network structures. In recent years, scholars have proposed various improvements on the basic methods, such as Fusion method of PCNN and multi-instance learning, PCNN and attention The fusion method of force mechanism, etc. Ji proposes to add the description information of the entity to the representation of the learning entity based on PCNN and Attention. The COTYPE model proposed by Ren and the residual network proposed by Huang enhance the relationship extraction effect[1].

In order to systematically review the relevant research results, we reviewed the review papers in recent years. It can be seen that the main advantage of the deep learning-based entity relationship extraction method compared with the classical extraction method is that the deep learning neural network model can automatically learn. Sentence features do not require complex feature engineering. Therefore, this paper focuses on deep learning to explore the method of entity relationship extraction[2].

2. Problem Definition and Solution Framework for Deep Learning Entity Relationship Extraction

2.1 Problem Definition

Entity relationship extraction, as an important task of information extraction, refers to extracting pre-defined entity relationships from unstructured texts based on entity identification. The relationship of entity pairs can be formalized as relational triples <e1, r, e2>, where e1 and e2 are entities, and r belongs to the target relationship set R{r1, r2, r3, ..., ri}. The task of relation extraction is to extract the relationship triplet from the natural language text. E1, r, e2>, thereby extracting text information.

The deep learning entity relationship extraction is mainly divided into two categories: supervised and remotely supervised. In the supervised, the method of solving the entity relationship extraction can be divided into two types: pipeline learning and joint learning: the pipeline learning method refers to the completion of entity recognition. Based on the direct extraction of the relationship between entities; the joint learning method is mainly based on the end-to-end model of the neural network, and at the same time complete the real identification and the extraction of relationships between entities. Compared with the extraction of supervised entity relations, the remote monitoring method is lacking. Manually labeling data sets, therefore, the remote monitoring method is more than one step of remotely aligning knowledge to the process of marking unlabeled data. The part of constructing the relational extraction model is not much different from the pipelined method of supervised domain. Based on deep learning Entity relationship extraction, entity relationship identification, and entity relationship classification are three concepts that are similar and related to each other. Specifically, relationship extraction deals with the same task in relation to relationship classification in its pipeline processing scenario. Refers to the entity pair directly in the case where the named entity pair in the sentence has been identified. Relationship classification; and relationship extraction in the joint learning scenario is to classify the relationship as a sub-task of its own. At this time, the relationship extraction refers to: the entity relationship extraction task is divided into two sub-tasks: named entity recognition and relationship classification. The joint learning model solves these two subtasks at the same time. The entity relationship recognition task is the same as the relationship extraction task[3]. In the actual processing, the semantic relationship between the entities is also discovered and recognized. Therefore, in some Chinese and foreign review literatures, the entity relationship extraction is sometimes called Identify entity relationships.

2.2 Problem Solving Framework

(1) Obtaining tagged data: a supervised method obtains a tagged dataset by manual tagging, and the remote monitoring method acquires a tagged dataset by automatically aligning the remote knowledge base;

(2) Constructing a word vector representation: there will be a tagged sentence segmentation, each word is encoded into a computer-acceptable word vector, and the relative position of each word to the entity pair in the sentence is obtained, as the position vector of the word, The word vector is combined with the position vector as the final vector representation of the word;

(3) Perform feature extraction: input the vector representation of each word in the sentence into the neural network, extract the sentence features using the neural network model, and train a feature extractor;

(4) Relationship classification: According to the pre-defined relationship types, the vector extracted by the feature is put into the nonlinear layer for classification, and the final entity pair relationship is extracted;

(5) Evaluation of classification performance: Finally, the relationship classification results are evaluated. The evaluation indicators and related data sets are detailed in Section 6.

3. Supervised Entity Relationship Extraction Method Based on Deep Learning

3.1 Supervised Entity Relationship Extraction Framework Evolution Process

The relationship extraction based on the supervised method in the deep learning method is a research hotspot of relationship extraction in recent years. It can solve the two main problems of artificial feature selection and feature extraction error propagation in the classical method, and combine the low-level features to form More abstract high-level features for finding Looking for the distributed feature representation of data. From the perspective of the neural network model based on supervised learning, the research mainly focuses on the fusion of multiple natural language features to improve the recognition accuracy.

3.2 Pipeline Method

The main process of relationship extraction based on the pipeline method can be described as: extracting the relationship of the sentence that has already marked the target entity pair, and finally outputting the triplet with the entity relationship as the prediction result. Some relationship extraction models based on the pipeline method are It is proposed that the network structure based on RNN, CNN, LSTM and its improved model has received a lot of attention from the academic community because of its high precision. The RNN has both internal feedback connections and feedforward connections between processing units. It can use its internal memory to process sequence information of arbitrary time series. It has the ability to learn the combined vector representation of various phrases and sentences of arbitrary length. It has been successfully applied in many NLP tasks. The method of relation extraction based on RNN model is Socher et al. first proposed in 2012 that this method assigns a vector and a matrix to each node in the analysis tree, where the vector captures the intrinsic meaning of the component, and the matrix captures how it changes the inclusion of adjacent words or phrases. Yi. This matrix vector RNN can learn the meaning of operators in propositional logic and natural language, and solves the problem that singlewordvector space models can not capture the meaning of long phrases, which hinders them from understanding the language more deeply. problem[4].

3.3 Joint Learning Method

The joint learning method directly obtains the entity triples of the existing relationship through the joint model of entity recognition and relationship classification. Because the objects modeled in the joint learning method are different, the joint learning method can be divided into parameter sharing method and sequence labeling method: parameters Sharing methods for entities and relationships Line modeling, while the sequence labeling method directly models the entity-relational triples.

Li et al. used this model in 2017 to extract the “Live-In” relationship between bacterial and bacterial locations, and made two improvements to the Miwa model based on practical applications: 1) To improve subtask identification from entities The error accumulation propagation problem that may be generated by the relationship classification subtask, a new relationship “Invalid_Entity” is introduced in the relationship classification subtask, and the entity generated in the entity identification subtask is verified to distinguish the effective entity from the invalid entity. Then classify the relationship between “Lives_In” and “not Lives_In” for the effective entity; 2) In the entity identification subtask, the method of gradually predicting the physical tag from left to right due to greed may bring errors between these tags. Propagation, that is, errors in previous predictions may cause new errors in subsequent predictions, so the original greedy search decoding in the model is replaced by beam search, because each step in the beam search can have multiple candidate predictions. If the best prediction is not correct, the candidate prediction can be selected according to the global score order, and the model is trained in the beam search with the early update technique to alleviate the entity label. Error propagation problem.

In 2016, Katiyar et al. first used the method of deep bidirectional LSTM sequence annotation to jointly extract viewpoint entities and IS-FROM, ISABOUT relationships. At the same time, it also proposed to add sentence-level constraints and relationship-level optimizations on the output layer to improve the model. Accuracy. However, this method can only identify the view entity and IS-FROM, IS-ABOUT relationship, can not extract the relationship between entities, the model can not be extended to extract other relationship types. After that, the model can not be extended for improvement. The problem, Katiyar et al., based on its 2016 model, first used the attention mechanism together with the two-way LSTM in 2017 to jointly extract entity and classification relationships. The model is shown in Figure 5. The entity recognition subtask and the relationship classification subtask share a coding layer representation (the coding layer includes the LSTM unit and the hidden layer). The model is in the entity identification subtask and Miwa et al. The human model is consistent, the entity recognition subtask is treated as a sequence labeling task, and the multi-layer dual LSTM network is used for entity detection. On the relationship classification subtask, the method improves the characteristics of Miwa et al. depending on part of speech tags and dependency trees. Disadvantages, entities based on entity recognition subtask output Sequence representation and shared coding layer representation, using attention model to classify the relationship; at the same time, the model can also extend and extract various defined relationship types, which is the first neural network joint extraction model in the true sense.

4. Conclusions

(1) Overlapping entity relationship identification at present, in the case of overlapping entity relationship identification, the existing entity relationship recognition model has not given a corresponding solution. Although the new labeling strategy proposed by Zheng solves the existence of parameter sharing method the problem of redundant entities really combines the two tasks into one sequence labeling problem, but the method still does not solve the overlapping entity relationship problem. Therefore, the overlapping entity relationship in the future will still be a major problem for scholars to study and overcome. In addition, Due to the introduction of Zheng's new labeling strategy, more improvements and developments can be made in this set of labeling strategies in the future to further improve the end-to-end relationship extraction task.

(2) Cross-sentence level relationship extraction. Nowadays, the relationship extraction task concentrates on classifying the pairs of entities identified in one sentence. According to the habit of natural language, the cases of entity pairs in different sentences are also very common. The existing referential elimination tasks can be used to refer to Generation object recognition and referencing object center word extraction effectively affect the performance of many natural language processing task systems, but its existence depends on the problem of strong artificial features and high precision. Therefore, fusion and improvement of the referential digestion and relation extraction model, It is a solution that can be studied in the future to solve cross-sentence level relationship extraction tasks.

(3) Relationship type OOV problem. Nowadays, in the mainstream methods of completing the relationship extraction task, the problem of out of vocabulary (OOV) is not effectively solved. For the relationship types that do not appear in the training set, the existing model framework cannot accurately predict the entity pair. The correct relationship type. In SemEval-2010's evaluation task 8, due to the consideration of the order of the entity pairs in the sentence instance, the other class was introduced to describe the instances that are not part of the existing relationship type, but this only reduces the existence relationship. The loss of the entity pair improves the ability of the model to judge the relationship, but the relationship between the entity pairs in the other class is difficult to define, the relationship is vague, and requires manual intervention and judgment. Therefore, the relationship type OOV problem is also a problem to be solved in the future. One.

Acknowledgement

This article was specially funded by Dalian University's 2019 Ph.D. Startup Fund (20182QL001) and 2019 Jinpu New District Science and Technology Project.

References

[1] Golshan PN,. (1995) A study of recent contributions on information extraction, New Technology of Library and Information Service ,8, 18−23

[2] Gan LX. (2016) Chinese entity relationship extraction based on syntactic and semantic features: IEEE Press, 8, 69−73.

[3] Ratnaweera A. (2004) Self-organizing hierarchical particle swarm optimizer with time-varying acceleration

[4] coefficients. IEEE Transactions on Evolutionary Computation, 6, 712-731.