Overview of Entity-Relationship Extraction

Zizi Zhang , Michael Smith

Tulane University, Tulane University
Zhangzizi9@tulane.edu,smithm22@tulane.edu

Abstract: Information extraction is a pivotal focus in the field of natural language processing, with entity-relationship extraction being a key component. This paper examines the historical development of entity-relationship extraction and discusses the characteristics of various entity-relationship extraction methods. Additionally, it presents an overview of several leading entity-relationship extraction techniques currently in use. The paper concludes with a forward-looking perspective on the future of entity-relationship extraction, particularly in the context of advancements in deep learning.

Keywords: Natural language processing; Entity relationship extraction; Deep learning.

1. History of the development of entity relationship extraction.

Entity relationship extraction was first proposed at the MUC conference in 1998, when it was mainly performed using lexicons and manual methods[1].The conference was commercially oriented, and the classification of relations and the annotation of entities on the English corpus was done by manual annotation, and the model was tested and evaluated to some extent.

Since then, the ACE conference has replaced the MUC conference and was merged into the TAC conference in 2009, which focuses on natural language processing and related applications[1] It has greatly contributed to the development and research of entity relationship extraction techniques.

After the ACE and MUC conferences, the natural language processing field started to focus on the SemEval conference. This conference focuses on connections between sentences, between utterances, etc.[3] . The conference defined the most common entity relations (cause-influence, product-producer, etc.), using lexicons and manual and traditional machine learning for the task of disambiguating English word meanings.

At the same time, with the development of the field of natural language processing, more and more experts have started to build up corpora, thus giving a great impetus to the development of entity-relationship extraction and substantially improving the performance of relational entity extraction.

2. Definition of entity relationship extraction

Entities in text are mainly nouns and specific words, while relations refer to the links between entities, such as syntactic links and syntactic links. However, entity-relationship extraction is the extraction of structured data from unstructured data text. Structured data is mainly described by entity-relationship triples, i.e. <e1, r, e2>, where e1 and e2 are entities and r are the relationship type. The extracted entity triples are stored in the database for easy access when building knowledge graphs and intelligent question and answer systems. Taking the sentence "Jiuzhaigou is located in Sichuan" as an example, the sentence is pre-processed to identify the two entities "Jiuzhaigou" and "Sichuan", and then "is located" is the relationship between the two entities.

3. Relationship extraction features

Relation extraction focuses on the analysis and processing of text, so relationships have 3 main characteristics.

The domain is too wide and the model is complex to build. Because the text domain is too wide, the model construction cannot achieve universality, and the same model extracts text from different domains with very different results and performance, so we need to build different entity relationship extraction models for different domains.

The data structure is diverse and there are 3 main types of data: structured data, semi-structured data and unstructured data. Structured data is mainly tables, semi-structured data is mainly logs, JSON documents, emails, etc. Unstructured data mainly includes text files, websites, social media, etc. Structured data is easy to manipulate, while semi-structured and unstructured data are cumbersome to manipulate.

The variety of relationships is complicated [3] The relationships contained in both Chinese and English texts are very complex and may be over- or under-sampled in the process of relationship extraction, which is also a test of the model.

4. Relationship extraction evaluation criteria

The results of relational extraction for the same domain are evaluated by accuracy, recall, and F1[4] . Where accuracy is the ratio of the number of correct predictions to the number of all samples; recall, also known as the check-all rate, refers to the proportion of correct predictions in all positive samples, i.e. how many of the positive samples the model found correctly; and F1 is the summed average of accuracy and recall. The formulae for accuracy, recall and F1 are as follows.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (2)$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recal}} \quad (3)$$

5. Main methods of entity relationship extraction

5.1. Based on manual annotation and semantic rules

Earlier entity relationship extraction was mainly based on

manual annotation and semantic rules. The rules for the structure of entities were defined in advance using linguistics, and then the processed utterance fragments were matched with the patterns to extract and classify the relations.

Aitken proposed rules for information extraction based on inductive logic programming techniques and natural language data, and completed tests on certain data where the value of F1 was as high as 66%. 2013 Han Hongqi et al. proposed a term hierarchical extraction method based on word rule template matching using the head and modifier features of compound terms, comparing the edge shared words existing between two terms ,constructing templates to determine the IS-A and PART-OF relationships between them, and the accuracy of their model could reach 92.5% [5] .

Rule-based and manually annotated relational extraction models are labour intensive and require the builder to be very knowledgeable about the domain. This approach has been successful in certain domains, but in other domains rule-based and manually annotated approaches are more expensive to use and have lower model performance.

5.2. Dictionary-driven relationship extraction based on

Based on the above problems, a lexicon-driven approach to entity relationship extraction has been developed, which matches entities in a given text by identifying strings and discriminating relationships by identifying verbs in the domain dictionary. This method is also labour intensive but it improves the accuracy of the extraction results and the value of F1 [6] The method is also labour intensive but it improves the accuracy of the extraction results and the value of F1 [8].

5.3. Relational extraction based on traditional machine learning

Traditional machine learning methods are based on language models and have achieved good results with a clear research direction. The methods are divided into 3 main categories respectively supervised, unsupervised and semi-supervised[7] .

Supervised learning studies the model from the training data and predicts the type of relationship for the test data. The text then needs to be processed in some way when it is fed into RE, and there are two main types of methods for processing text: the feature vector method and the kernel function method.

5.3.1. Feature vector based

A series of feature vectors are extracted mainly from contextual information, lexicality, syntax[8] , which are then classified by a classification algorithm such as

Naive Bayes, ME maximum entropy model

5.3.2. Based on kernel functions

The classification model is trained by calculating the similarity between two entities through a kernel function.

The use of supervised learning methods is limited by the corpus and is also not suitable for relational extraction in some open domains

5.3.3. Semi-supervised learning

Semi-supervised learning, also known as weakly supervised learning, uses the assumptions of the model to improve the generalization of the model to labelled samples under the condition that a small amount of data is labelled, with the unlabeled data being Corpus text.

5.3.4. Unsupervised learning - clustering

Both supervised and semi-supervised learning require the type of relationship to be determined in advance, yet in the presence of large amounts of data, we cannot predict all entity relationships in the data[9] . Several researchers have tried to solve this problem by basing on the idea of clustering. unsupervised learning was first proposed by Hasegawa et al. at the ACL conference in 2004, and most subsequent methods have improved on Hasegawa's work. The results show that clustering methods are very feasible in relation extraction.

First, they obtain news texts through crawlers and then start classifying articles according to their sources. Then, based on the semantic structure of the sentences, basic pattern clusters of entities that satisfy a set of constraints are extracted, and these entities are mapped according to the basic model to form sub-clusters so that each sub-cluster contains the same relationships between the entities.

The unsupervised approach needs to be based on a large-scale corpus. More relationship names are found in the data by training on a large amount of data. The method is not able to describe the names of the associations so the recall of the method is low.

5.4. Deep learning-based relationship extraction methods

With the rapid development of deep learning, more and more scholars put deep learning into the field of natural language processing, of which entity relationship extraction is the main embodiment. 2005 Che Wanxiang et al. proposed a feature vector-based machine learning algorithm to convert instances into numerical values and use the learned classification functions for entity relationship extraction[10] ; In 2016, Wan Changxuan et al. proposed Chinese entity relationship extraction based on syntactic-semantic features by combining the dependent syntax of each of two entities to obtain their combined features[11] ; in 2017, Liu, Kai et al. incorporated convolutional neural networks into entity relationship extraction by inputting vector feature matrices to convolutional neural networks for training classification models to achieve entity relationship extraction[12] ; Aone et al. proposed an end-to-end relationship and event extraction system, YangXiaoMing of Beijing University of Posts and Telecommunications proposed an enhanced data generation algorithm based on lexicon and instance intersection, and XuJin of University of Electronic Science [13] proposed a joint extraction model based on BERT and an improved multi-head selection mechanism.

The current deep learning-based entity relationship extraction is mainly divided into supervised and unsupervised, of which two supervised approaches are pipelined extraction and joint extraction.

Flowline extraction

Streamline extraction is to extract entities and relations separately, first extracting the entities from the text, then extracting the relations from the text, and finally matching the entities and relations. Early streamline extraction methods are mainly based on convolutional neural network and recurrent neural network structures. The early pipelined extraction method is mainly based on convolutional neural network and recurrent neural network structures. The pipelined extraction method is more frequently used, but it will produce error propagation and cause a certain amount of entity information redundancy.

Joint extraction

Joint extraction is the extraction of entities and relationships between entities at the same time. The main joint extraction models are parameter sharing based entity relationship extraction, sequence annotation-based entity relationship extraction and graph-based entity relationship extraction. The joint extraction model can reduce errors and avoid redundancy of entity information.

6. Future Trends in Entity Relationship Draws

At present, entity relationships are developing rapidly and extraction techniques are maturing, but they still require a great deal of effort from scholars to explore.

Improving the performance of entity relationship extraction models. Although the performance of the current extraction model is stable, the performance of the extraction model varies from domain to domain, so further optimisation of the model is needed to normalise the model.

The study of joint extraction models has been strengthened. From the above analysis, it is found that the drawbacks of pipeline extraction are too obvious, while joint extraction can precisely compensate for the drawbacks of pipeline extraction, but the development of joint extraction model is not particularly mature, and the performance of the model is not particularly stable, while in the process of extraction still consumes a lot of manpower, which motivates researchers to continuously optimise the performance of joint extraction model.

Improving the extraction dimensionality of the extraction model. Current extraction techniques are mainly aimed at extracting binary relations, but some texts have multiple relations, so if only binary relations are extracted there will be a lack of information.

7. Conclusion

In summary, entity relationship extraction has become an important research direction in the field of natural language processing, and its research has changed from requiring a lot of manual annotation to semi-automation based on deep learning extraction. With the rapid development of entity relationship extraction technology, it will have a positive impact on the construction of knowledge graphs and intelligent question and answer systems, so entity relationship extraction technology has a broad application prospect and significance.

References

[1] Liu Suwen,Cheng Wei,Qian Longhua,Zhou Guodong. Combining relation extraction with function detection for BEL statement extraction.[J]. Database : the journal of biological databases and curation, 2019, 2019.

[2] Xuefeng Wang,Ruopeng Yang,Yulong Feng,Dongsheng Li,Jianfeng Hou. A Military Named Entity Relation Extraction Approach Based on Deep Learning[P]. Algorithms, Computing and Artificial Intelligence, 2018.

[3] Shengbin Jia, Shijia E, Maozhen Li, Yang Xiang. Chinese Open Relation Extraction and Knowledge Base Establishment [J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),2018,17(3).

[4] Wang Chuandong, Xu Jiao & Zhang Yong, A review of entity relation extraction. Computer Engineering and Applications, 2020. 56(12).

[5] Han Hongqi et al. A term hierarchical relationship extraction method based on word form rule templates. Journal of Intelligence, 2013. 32(07).

[6] Xu Q, Research on Chinese named entity relationship extraction based on lexical semantic information, 2016, Taiyuan University of Technology.

[7] Wang Zheng, Zhu Lijun & Xu Shuo, A weakly supervised learning extraction method for entity relations. China Science and Technology Resource Guide, 2018. 50(02).

[8] Huang X et al. Feature combination-based relationship extraction for Chinese entities. Microelectronics and Computers, 2010. 27(04).

[9] Cao, Cady et al. A review of cluster analysis. Smart Health, 2016. 2(10).

[10] Che, Wan-Xiang, Liu, Ting & Li, Sheng, Automatic extraction of entity relations. Chinese Journal of Informatics, 2005(02).

[11] Gan, Lixin et al. Chinese entity relationship extraction based on syntactic semantic features. Computer Research and Development, 2016. 53(02).

[12] Liu K et al, Weakly supervised relation extraction for Chinese healthcare based on convolutional neural networks. Computer Science, 2017. 44(10).

[13] Xu Jin, Deep Learning-based Joint Extraction of Entity Relations, 2021, University of Electronic Science and Technology.