International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 05, Issue 02, May 2024
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Research on Relation Extraction Method Based on Active Learning

Duan Lianzhai
President University MIT S2
Bekasi, West Java, Indonesia
lark95@163.com

Abstract— The knowledge in contemporary society has exploded, and the most common knowledge is contained in unstructured natural language texts. Information Extraction technology expresses semantic knowledge in unstructured text through a set of mentioned entities, the relationships between these entities, and the events in which these entities participate. As a key part of information extraction, Relation Extraction technology provides important theoretical basis and use value for text knowledge understanding by judging the relationships between given entities. Currently, relationship extraction based on supervised learning requires a large number of labeled samples. Randomly selecting some data labels is not only a waste of data resources, but also directly affects the final accuracy of the classification model. In fact, with the development of data collection and storage technology, it has become very easy to obtain a large amount of unlabeled natural language text. Therefore, it is of great practical value to design an algorithm that can effectively utilize unlabeled sample sets for relationship extraction. In order to solve the above problems, this paper uses active learning as the starting point to implement a variety of sampling algorithms, mainly including uncertainty, diversity, representativeness and other algorithms. On the basis of verifying that active learning is suitable for relationship extraction tasks, through the fusion of multiple This sampling criterion ultimately yields an active learning sample selection strategy that is still effective under multiple data sets and multiple learning models. Experiments have proven that the multi-criteria fusion sampling strategy proposed in this article is an effective and universal strategy. Compared with multiple single-strategy sampling algorithms, it can achieve equivalent or higher classification accuracy on multiple data sets.

Keywords*: Active Learning, Deep Learning, Information Extraction, Relation Extraction*

I. INTRODUCTION

1.1 Background

In order to improve the quality of feedback from current search engines, knowledge maps organize the various types of information on the internet in a structured form to make it more accessible to humans, providing a way to, the ability to store and query all kinds of information on the internet.[1] When users search, Knowledge Atlas first analyzes the problem, semantics, and understands the real query needs of users, then queries the related knowledge in the ATLAS database and returns it to users, thus improving the search quality. [2]Knowledge mapping, which relies on big data and artificial intelligence, has become an infrastructure for managing knowledge on the Internet.[3]

In the process of constructing knowledge Atlas, information extraction is the most core technology, including Entity extraction, relation extraction, among which relation extraction is an indispensable part of information extraction, by judging whether there is a certain relationship between the given sentence entities, and further determining the relationship category, the text analysis is promoted from the language level to the content level.[4] The existence of some kind of semantic relationship between entity 1 and entity 2 is usually expressed as a triplet (entity 1, Relationship Class, entity 2), for example, in the sentence "The CPU is the core of the computer," the entity relationship is computer, Core Unit, CPU. [5]The different relations among the entities associate the independent entities to form the knowledge network, and the high-quality relation extraction can not only increase the scale of the knowledge map, but also guarantee the quality of the knowledge map, therefore, it is of theoretical significance and practical application value to explore and study the technology of relation extraction.[6]

In the traditional method of relation extraction, researchers need to design the semantic rules carefully by hand, according to the matching of different samples and different rules, the relationship between entities in the samples is given, but not only does this approach require the participation of domain-specific experts, it is also difficult to migrate to other areas, and therefore extremely costly?[7] With the success of deep neural network, the method of learning based on data representation is widely used in relation extraction, replacing the traditional method based on manual feature, kernel function, conditional random fields. Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and so on.[8] [9], [10]However, in order to get good performance, this kind of supervised learning needs a large number of mark-up samples. In order to avoid the time-consuming and labor-consuming problem of manual mark-up data, especially the huge amount of unstructured network data, consider using Active Learning techniques.[11]

Active learning aims at training an effective learning model with the lowest mark cost possible. [12]Through heuristic learning, the strategy actively selects the samples that are most helpful to the model to be labeled by human experts, and adds the labeled instances into the training set, iterative training was used to improve the generalization performance of the classifier. [13]With the exponential increase of all kinds of data in the information age, the problem of data marking has been paid more and more attention by the academic and industrial circles, significant advances have been made in theory and algorithms, and have been widely used in image processing, speech recognition etc.[14], [15]

This article mainly studies how to apply active learning technology in relationship extraction tasks. The meaning is that under the condition of small-scale labeled corpus, It can effectively utilize the potential information in large-scale unlabeled corpus to learn and select the most effective part of the corpus for manual annotation. In order to ensure that the relationship extraction model has a small labeling cost, achieve higher learning performance.[16], [17]

1.2 Research objectives and contents

The main objective of this work is to design a relation extraction method based on active learning, and to obtain a highly universal and transferable sample query strategy, compare it with random sampling on different sampling models and different data sets to verify its validity.

The main research contents of this paper include:

(1). A variety of active learning algorithms are implemented, including basic sampling methods based on uncertainty, representation and diversity, to verify the effectiveness of active learning algorithm in relation extraction task.

(2). On the basis of the basic sampling method, the sampling strategy is designed by fusing many kinds of standards. Finally, the optimal sampling strategy is obtained by contrast experiment on the relation sampling task.

(3). To implement a variety of relation extraction models, including CNN, BLSTM, R-BERT, etc. to verify the design of active learning, the algorithm is universal.[18]

II. RESEARCH METHODS

2.1 Relation extraction

As a link of knowledge map construction, relation extraction is used to extract entity relationship triples from unstructured text to form structured knowledge. The traditional relation extraction based on manual design rules, which relies heavily on the features of manual design and the quality of extracted features, has great limitations, with the success of deep learning in the fields of image, speech and so on, many researches on relation extraction have also introduced neural networks to extract features of sentences automatically, which not only reduces the need for feature engineering, but also reduces the need for energy, and it can achieve good extraction results.

Deep learning-based extraction models treat relational extraction as a Multi classification problem, and the model framework is shown in figure 2.1:

Entity relationship extraction framework based on deep learning	5. Evaluate performance: Precision, Recall, F1 score
	4. Relationship classification: Softmax
	3. Feature extraction: CNN, RNN, Bi-LSTM
	2. Construct word vectors: Word vectors, Position vectors
	1. Access to tagged data: Manual marking

Figure 2.1 entity-relationship extraction framework based on deep learning

(1). Access to tagged data: access to tagged data sets through Manual marking.

(2). Construct word vectors: by segmentation of the marked text and mapping it to the corresponding word vector.

(3). Feature extraction: sentence vectors composed of word vectors are fed into a supervised classifier to extract sentence features.

(4). Relationship classification: after the linear/nonlinear change of the sentence feature vector, it is sent into Soft max classification to get the object and entity relations.

(5). Evaluate performance: Relationship Classification results were assessed by indicators such as F 1 scores.

In relation extraction using convolutional neural network, the model is first embedded by pre-trained or randomly initialized words (Word Embedding) expresses the Word in the sentence as the Word vector, by splicing the entity Word vector in the sentence and its upper and lower, the relative position of the text represents the position vector of the entity Word to get the final Word vector representation, then, a CNN network is used to extract the sentence, and the level of the feature C1,C2, ...Cn, and then the relationship categories of the sentences are obtained by pooling and full-connection layers.

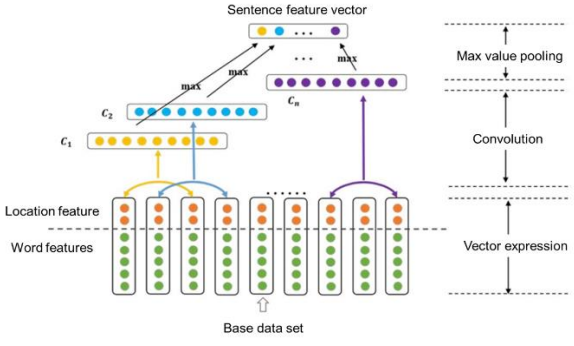


Figure 2.2 CNN-based relation extraction model

Based on the attention mechanism, the BLSTM neural network model develops the relation extraction task. First, each word is mapped to the low-dimensional space by Embedding layer, and then the bidirectional LSTM obtains the high-level features from it, and then we multiply that by the weight vector, generated at the attention level, so that the word-layer features in each iteration are merged into sentence-layer features, and finally we use the sentence-level feature vector for relational classification. The model structure is shown in figure 2.3.

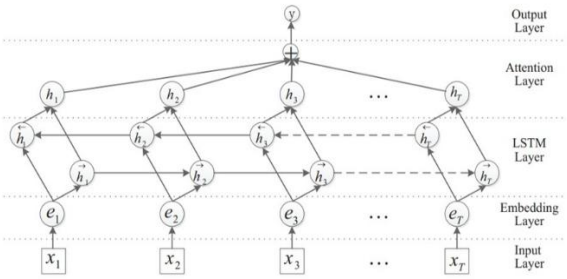


Figure 2.3 Att-BLSTM-based relational extraction model

BERT is a pre-trained bi-directional language model with Transformer as the feature encoder, and Transformer is a deep network superimposed on the self-attention mechanism, which is not only capable of capturing long-range features, but also has good parallel computing ability, using the R-BERT model to deal with the task of Chinese language relations. Using the pre-trained BERT language model, we combine the information from the target entities with the information from the BERT language model, add identifiers before and after the entities according to the input requirements of the BERT to indicate the location of the entities, merge the input sentences and entity pairs of information into an input sequence, and output the final implied state vectors of the identifiers and the final implied state vectors of the two target entities, and then classify them through the Softmax layer by combining the vector information of the three parts of the three parts of the vector through the linear/nonlinear change. The model structure is shown in Figure 2.4.

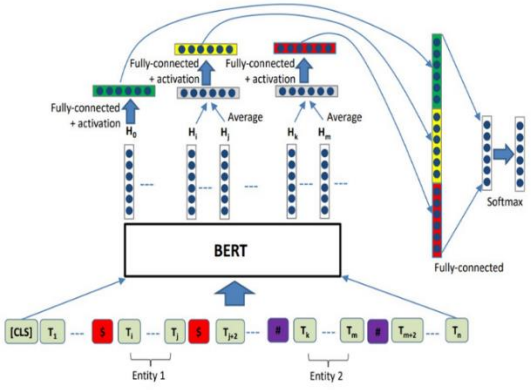


Figure 2.4 BERT-based relation extraction model

2.2 Active learning

2.2.1 Active learning algorithm model

Active learning through the design of a reasonable sampling method, from the unlabeled samples to select the most help the basic model to obtain better performance of the samples after marking, adding the labeled training set, re-training the basic model, iterate until the model meets certain performance requirements or exceeds the mark-up cost. At the heart of active learning is the learning engine and the sampling engine.

Learning engine refers to the basic model, that is, the classifier is trained on the labeled data set, and on the test set, to verify the generalization performance. The sampling engine is to select the unlabeled samples on the unlabeled data set by using the instance selection algorithm. These samples are labeled by human experts and used by the learning engine. The whole learning process is the learning engine, and the sampling engine work iteratively, and finally get a good enough performance classifier at an acceptable mark cost. Figure 2.5 shows the active learning model.

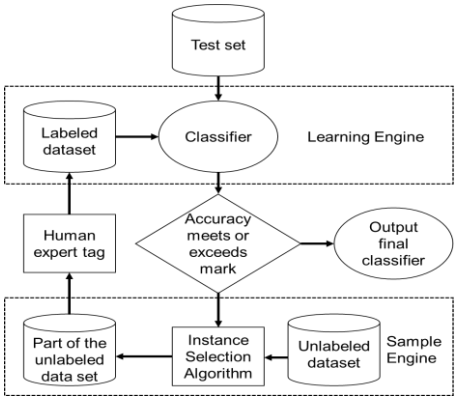


Figure 2.5 diagram of the active learning model

2.2.2 Introduction to mainstream active learning

According to the different ways of active learning to select unlabeled samples in different application scenarios, the active learning algorithm is divided into three types:

Query synthesis algorithm Stream-based algorithm Pool-based algorithm

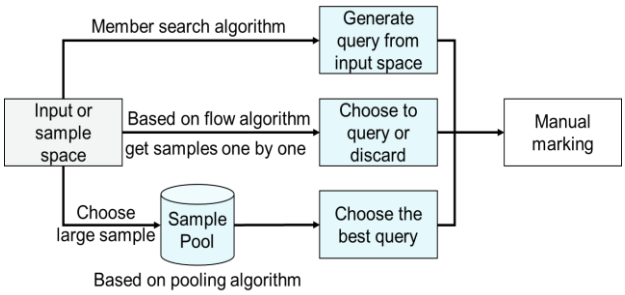


Figure 2.6 three active learning algorithms

The query synthesis algorithm is the first algorithm that learns by querying samples, by asking an expert for the most helpful sample tokens in the entire input space. The query synthesis algorithm is very effective in a restricted domain, but querying all the samples in the input space without considering the actual distribution of the samples leads to a lot of work that is synthesized but not meaningful. For example, when the sample data is in the form of text and the underlying model is Generative Adversarial Networks (GAN), the algorithm may create text that is similar to the lexicon of a normal utterance, but has no actual semantic information, and it is clearly pointless to leave these data samples to be annotated by experts. Therefore, this kind of algorithm is not suitable for application scenarios where the samples are labeled by human experts.

In order to solve the above problem, researchers have proposed a stream-based sampling strategy, in which all unlabeled samples falling in the sample space are sequentially labeled or discarded according to the sampling strategy. Generally speaking, this sampling strategy needs to compare the information content of the unlabeled samples one by one with a predefined fixed threshold, so the overall structure distribution of the unlabeled samples and the differences between samples cannot be obtained. It is only applicable to intrusion detection and information acquisition scenarios.

To address this shortcoming, the researchers propose a pool-based sampling strategy. Considering all unlabeled samples as a "pool". selectively labeling samples from the pool. Compared with the stream-based algorithm, by calculating the information content of all unlabeled samples in the pool and selecting the ones with the best information content to be labeled, this strategy avoids the need to set a fixed threshold, and the need to select the samples with the best information content. It avoids setting a fixed threshold and avoiding the situation of setting a fixed threshold and querying for meaningless samples. Therefore, it has become the most widely researched algorithm in the field of active learning. The algorithms have been applied in video retrieval, text categorization, information extraction, and so on.

III. RELATION EXTRACTION BASED ON ACTIVE LEARNING

3.1 Extraction model

In order to get the relation categories between entities, it is often necessary to combine the relation between the target entities and the semantic information of sentences effectively, relation classification. Bert is divided into two stages: pre-training and fine-tuning. In the pre-training stage, Transformer is used as a feature encoder to perform two tasks, namely, masking language model (MLM) and next sentence prediction (NSP), the local and global feature representations of the sequences are obtained, and the network structure of Bert is shown in figure 3.1.

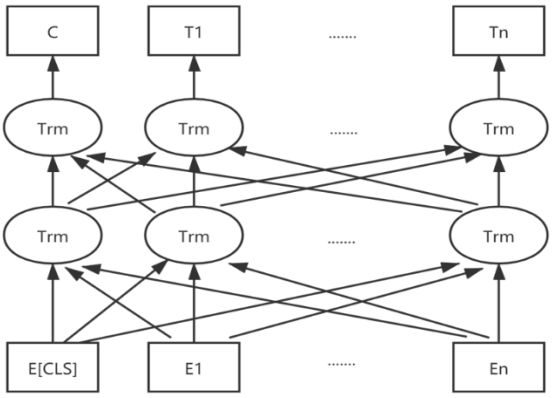


Figure 3.1 Bert model structure

As you can see, that BERT is essentially a stack of Transformer encoders (Trm module in the figure), which is originated from the Attention mechanism, and it completely abandons the traditional RNN, and the whole network structure is composed of the Attention mechanism. The features are extracted by the Transformer encoder, which is a popular feature extractor because it not only has good parallel computation capability for fast training, but also can capture the deeper connections between utterances.

The input vector of BERT consists of three parts: word vector, position vector, segment vector, and, in order to facilitate the subsequent fine-tuning phase of the classification task, a [CLS] token is added at the beginning of each input sequence. 15% of the tokens are randomly masked as the training samples for the MLM task, and among these samples, 80% are replaced by mask tokens, 10% by In these samples, 80% are replaced by a mask token, 10% are replaced by a random token, 10% are kept unchanged, and then the encoder predicts these tokens according to the context. Under such a task-driven approach, through iterative training, the BERT model can learn the syntactic, grammatical, and contextual features of each token very well.

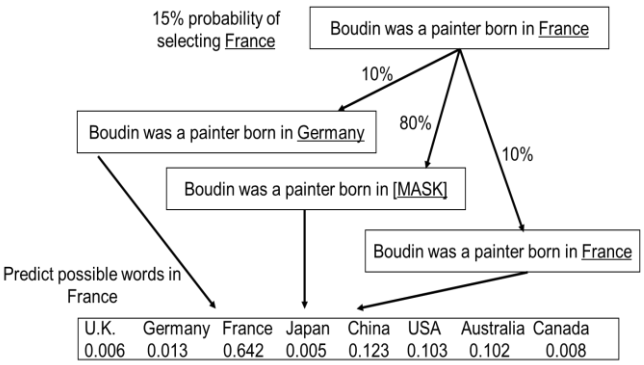


Figure 3.2 MLM task execution

Since the [CLS] token does not participate in masking in the pre-training stage, the position faces all positions in the whole sequence to do Attention, so that the output of the [CLS] position is sufficient to express the information of the whole sentence, while the vectors corresponding to other tokens pay more attention to the semantic syntax and contextual information expression of the token, so far, after the pre-training, we get a generalized performance of Thus, after pre-training, a language model with sufficiently good generalization performance is obtained that can characterize the information of the corpus.

In the fine-tuning stage, for the relationship extraction task, in order to enable R-BERT to locate the position of the two entities, the special character "$" is added before and after the first entity and the special character "#" is added before and after the second entity, and three parts of the features are utilized for the final classification of the relationship, including the [CLS] final implied state vector, and the implied state vectors of the two entities. The specific process is as follows: notate the output vector of [CLS] position as H_0, the individual word vectors of entity $e1$ are H_i to H_k, and the individual word vectors of entity $e2$ are H_j to H_m and the output vectors of the two entities are denoted as:

$$H_1 = \frac{1}{k-i+1}\sum_{t=i}^{k} H_t \tag{3.1}$$

$$H_2 = \frac{1}{m-j+1}\sum_{t=j}^{m} H_t \tag{3.2}$$

Where $H_1, H_2 \in R_{n\times d}$, n is the batch size, and d is the size of the hidden state of BERT. The two entity vectors and the [CLS] positional input vectors are nonlinearly activated (tanh) and then passed through the fully connected layer to obtain H'_0, H'_1, H'_2,

$$H_0^{'} = W_0[tanh(H_0)] + b_0 \tag{3.3}$$

$$H_1^{'} = W_1[tanh(H_1)] + b_1 \tag{3.4}$$

$$H_2^{'} = W_2[tanh(H_2)] + b_2 \tag{3.5}$$

Where $b'_{0\text{-}2} \in R_{d\times d}$, $W'_{0\text{-}2} \in R_{d\times d}$, $H'_{0\text{-}2} \in R_{n\times d}$, $W_1 = W_2$, and $b_1 = b_2$. i.e., $W_1$and W_2. b_1 and b_2 share parameters. By splicing these three feature vectors and inputting them into the fully connected layer, finally soft max classification is used to obtain the relation

$$h^{'} = W_3[concat(H_0^{'}, H_1^{'}, H_2^{'})] + b_3 \tag{3.6}$$

$$p = softmax(h^{'}) \tag{3.7}$$

Where $W_3 \in R_{l\times 3d}$, h' $\in R_{n\times l}$, l is the number of relational categories, and p is the predicted relational category.

3.2 Basic sampling method

This paper mainly focuses on the pool-based sample sampling strategy as the object of research, to study the development of a sample sampling strategy suitable for the relational extraction task, to ensure that the model achieves a certain performance while reducing the labeling cost as much as possible, i.e., maintaining a pool of unlabeled samples. By actively learning the sample sampling strategy, we iteratively select the labeled samples and train the model, so that the generalization ability of the model can be rapidly improved. The selection strategy generally follows the greedy idea, i.e., each iteration selects the sample with the largest (or smallest) attribute from the unlabeled sample set to be labeled.

3.2.1 Uncertainty-based sampling methods

The main idea of the selection strategy of the uncertainty-based sample sampling method is to select the samples from the unlabeled sample set that give lower confidence to the classification model, and by comparing the confidence size of the samples in the model classification results, determine the information content that each sample to be selected can bring to the classifier, and select the sample that can bring the most information from the unlabeled sample set to obtain the labeling to be added to the labeled sample set. Specifically for the entity-relationship extraction task, since an utterance has multiple possible labels, the uncertainty of a sample can be measured by the confidence with which the relationship is predicted to be in each category. last confident algorithm takes the most probable category for each sample as its representative category and selects the sample with the highest uncertainty (i.e., the lowest confidence) based on the corresponding confidence of the representative category:

$$x_L^* = \arg\max_{x\in U} 1 - P_\theta(\hat{y}|x) \tag{3.8}$$

Where:

$$\hat{y} = \arg\max_{y} P_\theta(y|x) \tag{3.9}$$

However, the above method only considers the class with the highest a posteriori probability and simply ignores the other classes, which is likely to have metric errors. In the actual classification results, the highest two categories with the highest confidence level are often close to the predicted probability, in this case, the above minimum confidence strategy is improved, margin sampling uses the difference between the confidence level of the largest and the next largest category in the prediction results of each sample as a measure of the sample uncertainty, and it is obvious that the smaller the difference between the confidence level of the sample, the more difficult it is to distinguish between the actual categories, therefore, we can choose this category to obtain the labeling to bring more

effective information to the base model. Selecting this category of samples for labeling can bring more effective information to the base model, and the sampling strategy is shown below:

$$x_M^* = \arg\min_{x \in U} \{P_\theta(\hat{y}_f|x) - P_\theta(\hat{y}_s|x)\} \quad (3.10)$$

Where：

$$\hat{y}_f = \arg\max_y P_\theta(y|x) \quad (3.11)$$

$$\hat{y}_s = \arg\max_{y, y \neq \hat{y}_f} P_\theta(y|x) \quad (3.12)$$

The closer the predicted probabilities of the two most likely categories obtained by the model's predictions, the more difficult it is for the model to determine their categories, and therefore the more deserving they are of being labeled. For Mult categorical data, considering only two categories ignores a great deal of information. From the point of view of using all the classification results of the sample and the corresponding probability to calculate the value of the sample, entropy sampling calculates the information entropy of all the classification results of the sample by introducing the method of information entropy, it is obvious that the samples with larger information entropy can bring more changes (amount of information) to the classification model, therefore, the samples with larger information entropy should be prioritized to be labeled, as shown below:

$$x_E^* = \arg\max_{x \in U} -\sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

Where y_i covers all possible labels so that the overall probabilistic information of the sample can be captured and a more accurate amount of information can be obtained.

3.2.2 Diversity-based sampling methods

Uncertainty-based sampling methods only consider the problem of the amount of information in a single sample and ignore the problem of information redundancy in the selected samples, so the sampling strategy can also be considered in terms of the diversity of the samples. If an unlabeled sample is too close to the samples in the labeled samples, then it means that it has a lot of similar information to those labeled samples it is close to and has no labeling value.

Therefore, a diversity-based sampling approach means prioritizing the unlabeled samples that are least similar to all the samples in the labeled sample set, and adding them to the labeled dataset will make the distribution of samples in that set as spread out as possible. Commonly used similarity criteria are Euclidean distance, Pearson's correlation coefficient, cosine distance, etc. For the relational extraction task, the sentence vectors are obtained by averaging the word vectors from the sentence meanings, and the similarity between two sentences is measured by the cosine distance between the two sentence vectors.

$$x_D^* = \arg\min_{x_u \in U} \sum_{x_l \in L} CosineDis(x_u, x_l) \quad (3.14)$$

Where:

$$CosineDis(x_u, x_l) = \frac{x_u * x_l}{|x_u||x_l|} \quad (3.15)$$

3.2.3 Representative-based sampling method

The representative-based sampling method considers the overall data distribution in the unlabeled dataset and selects the most representative samples that can better represent the sample space, in order to improve the differentiation ability of the underlying model, and ultimately achieve the purpose of improving the efficiency of the active learning algorithm. Taking Figure 3.3 as an example, the straight line in the figure represents the decision boundary, the squares and triangles represent the two types of labeled samples, and the circles represent the unlabeled samples. Because sample A is located on the decision boundary, it has the highest uncertainty, but in fact sample B will provide more effective information to the base model, this is because sample A belongs to an isolated point in the sample distribution, with low information density, while sample B has the commonality of the nearby unlabeled samples to some extent.

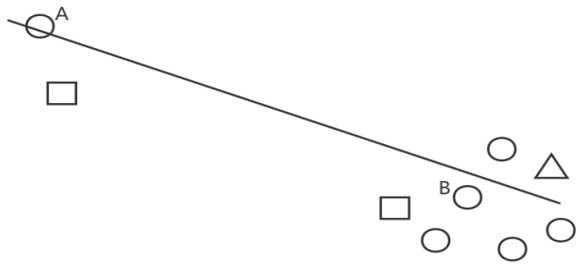


Figure 3.3 diagram of the active learning model

The specific operation is as follows: all the samples in the unlabeled sample set are clustered and divided into multiple class clusters, so that the difference between the samples within the same class cluster is as large as possible, and the difference between the samples between different class clusters is as small as possible, and then the samples with the largest information density are selected from them, i.e., the samples that are closest to the center of the class clusters as the samples of the selection, and their sampling formulas are as follows.

$$x_R^* = \arg\min_{x \in UC} EuclidDis(x, c) \quad (3.16)$$

$$EuclidDis(x, c) = \sqrt{\sum_{k=1}^{m} (x_k - c_k)^2} \quad (3.17)$$

Where UC is the unlabeled sample cluster, x_k represents the k feature of the sample x after the base model feature extraction, and ck represents the k feature in the center of the cluster of this class. In most cases, the efficient K-means algorithm is used for clustering, and homogeneity and completeness are used to measure the clustering effect.

3.3 Integrated Sampling Strategies

The methods mentioned in the previous section all use a single selection criterion to select samples, however, a single criterion is difficult to guarantee that it works on all underlying models and datasets, and thus how to effectively integrate multiple criteria to select samples becomes the focus of our research. There are two ways of combining multiple criteria considered. One is to score each sample by giving different weights to different sampling criteria, and those samples with the highest total scores are the target samples. The second is to use different sampling criteria layer by layer, the screening results of the previous layer are used as the candidate samples of the next layer, and the candidate samples are continuously subdivided, so that the samples that remain are the ones that are most capable of taking multiple criteria into account.3.3.1 a multi-criteria-based weighted sampling strategy.

3.3.1 multi-criteria based empowered sampling strategy

Since both diversity and representativeness sampling methods are used to avoid selecting samples with redundant information so that the set of labeled samples conforms to the full sample space as much as possible, these two components are combined with equal weights, and uncertainty-based sampling methods are given high weights considering that they are from the perspective of the amount of information in the samples, which has been proved to be powerfully effective in various tasks in the past. In order to effectively balance the effective dispersion of data distribution among the samples and the high information content within the samples, the uncertainty, diversity, and representativeness-based sampling method is finally given a weight of 2:1:1 to weight the unlabeled samples according to their respective criteria, and the higher the total score, the more worthy the samples are to be labeled, and the algorithm is described as Algorithm 1.

Algorithm 1 multi-criteria assignment sampling strateg:

Input:
Labeled sample set L
Unlabeled sample set UU
Sampling engine SE
Learning engine LE
Process:
1: $Train\ (LE, L)$

2: repeat

3: $L_f = Select\ (L)$ by (3.13)

4: $L_d = Select\ (L)$ by (3.14), (3.16)

5: $WeightedScore(L_f\ , L_d)$

6: for k = 1: m do

7: $x*$ = arg max $Score(x)$

$$x \in U$$

8: $Label\ (x*)$

9: $L \leftarrow L \cup \{x*\}$

10: $U \leftarrow U \setminus \{x*\}$

11: end for

12: $Train\ (LE, L)$

13: $Test\ (LE)$

14: until accuracy meets or exceeds the marked cost

15: return LE

The multi-criteria-based empowerment sampling strategy effectively combines the three sampling methods of uncertainty, diversity and representativeness by assigning different weights, and is measured by a scoring mechanism, so that the samples with the highest scores in the final pool of unlabeled samples are the ones we need to label.

3.3.2 multi-criteria-based layer-by-layer sampling strategy

In the introduction of the representative-based sampling method in the previous section, we choose the sample closest to the center of the class cluster as the representative sample of the class cluster, but in fact, the sample at the center of the cluster is not necessarily a good representative of the overall sample of the class cluster, inspired by the diversity-based sampling method, in this method, we believe that the sample that has a high degree of similarity with all other samples within the class cluster is the sample that is most representative of the class cluster of the overall samples. Considering that uncertainty-based sampling is biased towards picking samples closest to the decision boundary, these samples can lead to faster model convergence and uncertainty-based sampling tends to provide good performance gains across tasks. Therefore, by calculating the uncertainty of all candidate samples, some samples with the largest uncertainty are selected as the initial screening sample pool, and immediately after that, the initial screening sample pool is clustered, and the

entropy value of the similarity of the sample sentence vectors is calculated within each class cluster, and the samples with the largest entropy value are selected as the representative samples of the class cluster to be labeled. Algorithm 2 is described as follows:

Algorithm 2 multi-criteria layer-by-layer sampling strategy：

Input:

Labeled sample set L

Unlabeled sample set UU

Sampling engine SE

Learning engine LE

Process:

1: $Train\ (LE, L)$

2: repeat

3: $L\ f = Select\ (L)$ by (3.8), (3.10), (3.13)

4: $Call = Cluster\ (L\ f)$

5: for $C \in Call$ do

6: $x_c^* = \arg\max_{x_{cur} \in C} [\sum_{x \in C, x \neq x_{cur}} CosineDis(x_{cur}, x)]$

7: $Label(x_c^*)$

8: $L \leftarrow L \cup \{x_c^*\}$

9: $U \leftarrow U \setminus \{x_c^*\}$

10: end for

11: $Train\ (LE, L)$

12: $Test\ (LE)$

13: until accuracy meets or exceeds the marked cost

14: return LE

While each sampling criterion is measured as much as possible by means of multilayer selection, the computational overhead of the algorithm does not increase much compared to the basic sampling strategy because the inputs of each layer are the set of samples "filtered" by the previous layer.

IV. EXPERIMENT AND ANALYSIS

4.1 Experiment data

This experiment has three objectives, the first is to verify that active learning can effectively reduce the sample labeling cost in a relational extraction task, and the second is to verify that the proposed multi-criteria sampling-based algorithm can achieve superior or comparable performance compared to the base active learning algorithm. The third is to verify that the proposed active learning algorithm is model- and task-independent or can be cross-task and cross-model.

First, on each dataset is divided into a training set and a test set, and then the training set is divided into a set containing a small amount of labeled data and a set containing a large amount of unlabeled data.

Second, multiple basic sampling strategies are implemented and compared with random sampling under the same conditions to compare the performance of the model on the training sets selected by different strategies. In each iteration, the algorithm selects samples from the unlabeled set for labeling query according to different sampling strategies, removes the labeled data obtained from the query from the unlabeled set and adds it to the labeled set, and then re-trains the learning model based on the labeled set and evaluates the model on the test set.

The experimental data uses the open-source Information Extraction dataset of Baidu Brain, which includes 50 relational categories to be pre-processed, including the replacement of illegal characters in the collected text dataset, deletion of duplicated samples, and the remaining 200,000 samples, and then randomly samples from it with 5 samples of size 12000, which contains 4000 initial labeled samples for the training set and 2000 labeled samples for the test set. Then five datasets with sample size of 12000 are randomly sampled from the dataset, which contains 4000 initial labeled samples, 6000 unlabeled samples in the training set, and 2000 labeled test samples in the test set. In the active learning training phase, 200 samples from all unlabeled samples are labeled in each iteration, and the set is updated, with a total of 20 iterations.

4.2 Experimental Evaluation Metrics

In the relationship extraction task, the F1 score is mainly used as the evaluation criterion, which can effectively measure the model performance through the weighted sum of accuracy and recall.

$$F_1 = 2 \times \frac{precision * recall}{precision + recall} \quad (4.1)$$

Active learning aims to minimize the tagging cost associated with corpus tagging, and a common active learning evaluation metric is the percentage reduction in the number of labeled samples for this sampling strategy compared to random sampling when the base model achieves optimal performance, with the following formula:

$$P = \frac{N_{random} - N_{active}}{N_{random}} \times 100\% \quad (4.2)$$

In Eq. 4.2 $Nrandom$ is the minimum number of samples required for random sampling to achieve optimal performance, $Nacti\ ve$ is the minimum number of samples required for the active learning strategy to achieve optimal performance. However, in practice, the model needs to be trained on the basis of all samples being fully labeled to get the best performance, but if the query needs to be queried to label all the samples, it violates the main idea of active learning, i.e., it fails to achieve the purpose of reducing the labeling cost. Therefore, by assuming that the number of queries is fixed and small, the improvement in the performance that the model can achieve under the active learning strategy compared to random sampling is used as a measure of this algorithm, which is a better indicator of the effectiveness of active learning, as shown in the following equation:

$$P = \frac{P_{active} - P_{random}}{P_{random}} \times 100\% \tag{4.3}$$

In Eq. 4.3 $Pactive$ represents the model performance after a few active learning strategy queries, $Prand\ om$ represents the model performance after a few randomly sampled queries. By observing the model performance after a small number of samples can quickly get the difference between the advantages and disadvantages of different sampling methods.

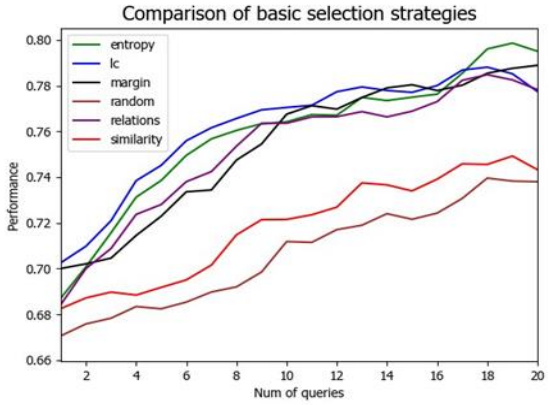
4.3 Experimental Results and Analysis

In order to verify the practical application effect of the various strategies based on active learning proposed in this chapter, three groups of comparison experiments are set up in the experimental session, and each group of experiments is set up with a control experiment to compare the accuracy curves of different active learning algorithms in the query process.

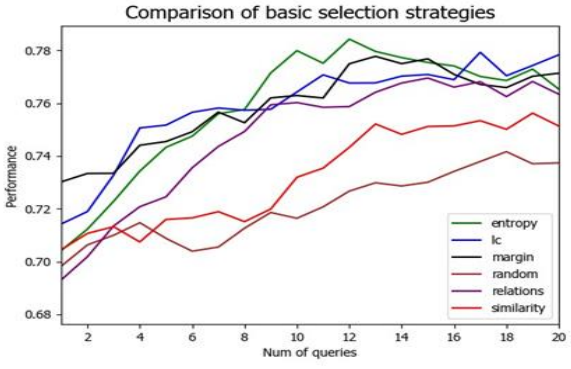
Experimental group 1: For the five different basic sampling algorithms, random samples were taken as the benchmark control to set up the comparison experiments, and the experimental results are shown in Fig. 4.1.

Among them, (a)-(e) represent the learning curves of multiple basic sampling strategies on five different datasets, where entropy, lc, and margin represent the entropy sampling algorithm, last confident algorithm, and margin sampling algorithm, respectively, in the uncertainty-based sampling method, relations represents representative-based sampling method, similarity represents diversity-based sampling method, and random represents random sampling. The horizontal coordinate represents the number of queries, and the vertical coordinate represents the performance improvement based on the R-BERT base model, at the same number of queries, the better performance indicates a better sampling strategy.

On datasets (a), (b), (c), all active learning algorithms significantly outperform random sampling, among which the three indeterminacy-based sampling methods are the most prominent, which verifies that the indeterminacy-based sampling methods are not only effective in other categorization tasks, but also applicable to the relational extraction task. Among them, the entropy sampling algorithm utilizes all possible classification information of the sample from the perspective of information entropy, which is the most effective. On datasets (d), (e), although the sampling method based on diversity and representativeness is worse than random sampling in the first few rounds of query iterations, it rises strongly in the subsequent performances and still outperforms random sampling in the end, verifying that unnecessary labeling costs can be saved by avoiding information redundancy among samples. Combining the analysis results obtained from multiple datasets, it can be verified that the active learning strategy is fruitful in the relational extraction task, where the uncertainty boosting performance is the best, among which the uncertainty-based entropy sampling algorithm, which integrally takes into account the probability distributions of all the categories, performs the best.

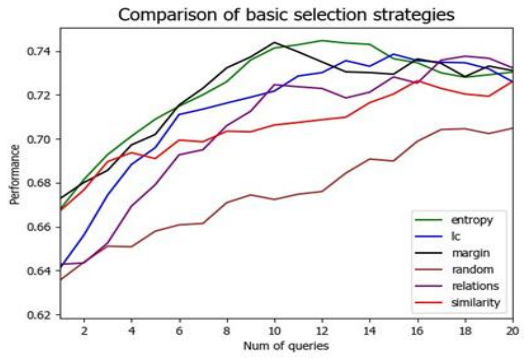


(a)

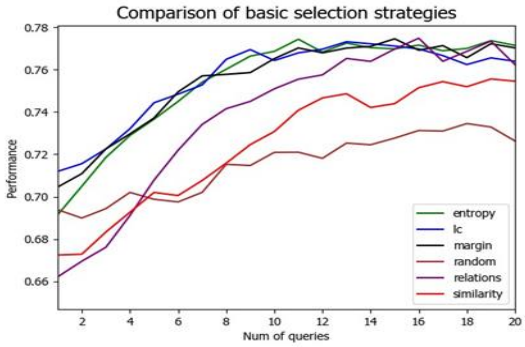


(b)

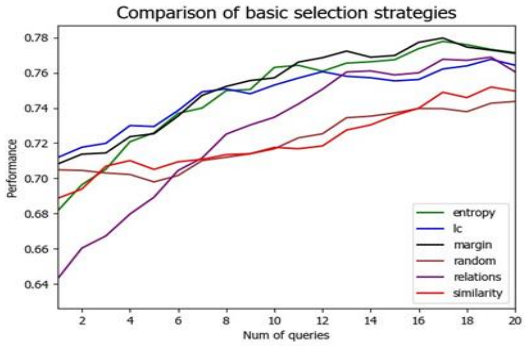
International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 05, Issue 02, May 2024
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index



(c)



(d)



(e)

Fig. 4.1 comparison of basic sampling strategies on different corpora

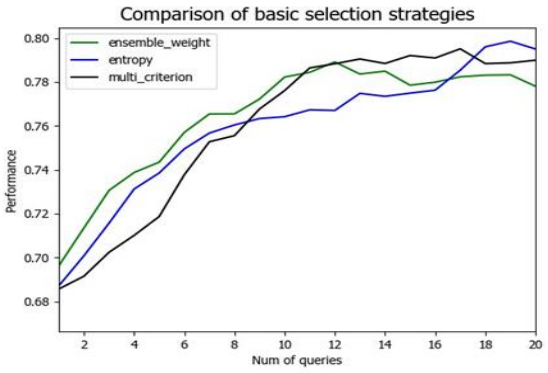
The sampling method of sex and random represent random sampling. The horizontal coordinate represents the number of queries, and the vertical coordinate represents the performance improvement based on the R-bert model.

On data sets (a), (b), and (c), all active learning algorithms significantly outperformed random sampling, with the three sampling methods based on uncertainty being the most prominent, uncertainty-based sampling methods were validated not only in other classification tasks

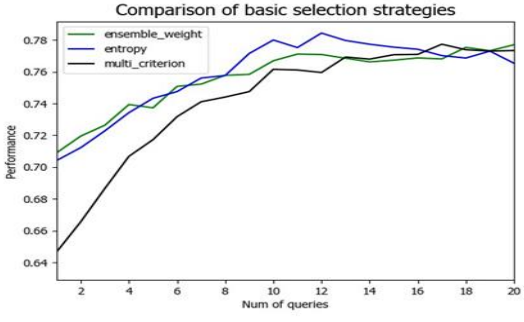
It is also applicable to the relation extraction task. Among them, entropy sampling algorithm from the perspective of information entropy, the use of samples, all possible classification information, the best effect. On datasets (d) and (e), although diversity-and representativeness-based sampling was worse than random sampling in the previous rounds of query iterations, it rose strongly in subsequent performance, and the final result was still better than random sampling, it is verified that unnecessary mark-up costs can be avoided by avoiding information redundancy between samples. The analysis results obtained from the data sets can verify that the active learning strategy is effective in the relation extraction task, and the uncertain promotion performance is the best in the relation extraction task, among them, entropy sampling algorithm based on uncertainty, which considers all kinds of probability distribution, performs best.

Experimental Group 2: For the 2 different integration strategies, the best performing basic sampling strategy is used as a benchmark control to set up a comparison experiment, and the experimental results are shown in Figure 4.2.

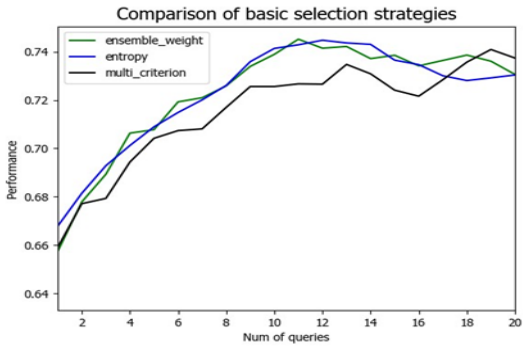
Among them, ensemble weight represents the multi-criterion-based empowerment sampling strategy, multi_criterion represents the multi-criterion-based layer-by-layer sampling strategy, and entropy represents the entropy sampling method in the best uncertainty-based sampling method in the basic sampling method. The horizontal and vertical coordinates represent the same meaning as experimental group I, by comparing which strategy performs better when the number of queries is the same.



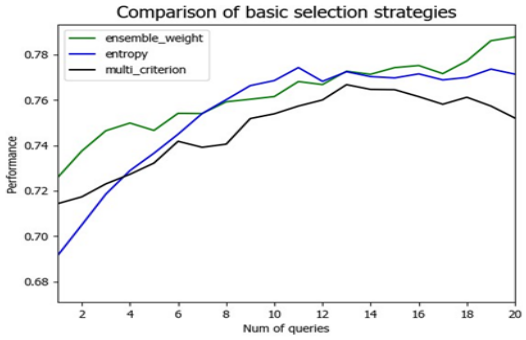
(a)



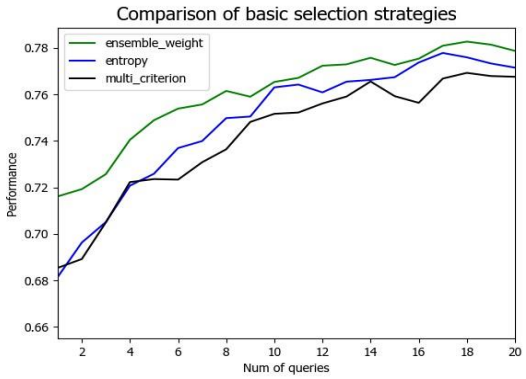
(b)



(c)



(d)



(e)

Fig. 4.2 comparison results of multiple-standard sampling strategies on different corpora

In datasets (a), (d), (e), a multi-criteria-based weighted sampling strategy performed better than entropy sampling, and in datasets (b), (c), it performed nearly as well as entropy sampling. By 2:1:1 weighting synthesis, the strategy of measuring multiple sampling standards not only considers the effective dispersion of data distribution among samples, but also considers the high information content within samples, it can perform better than single-base sampling strategy on many data sets, so it is proved to be effective. However, the performance of the multi-standard layer-by-layer sampling strategy is slightly worse than entropy sampling in most data sets, which shows that not all multi-standard strategies can be effective, and some integration strategies do not even have a single standard, good.

Experimental Group3: Compare the performance of three different base models on the same dataset on the single best basic sampling algorithm, multi-criteria based empowered sampling strategy, and random sampling to verify the model irrelevance. The comparison results are shown in Table 4.2.

Table 4.2 Experimental results of different models

Basic Models	Random Sampling	Optimal Basic Sampling Strategy	Multi-criteria sampling strategy
CNN	0.4312	15.38%	18.51%
Att-BLSTM	0.5574	8.45%	11.76%
R-BERT	0.7305	6.31%	7.68%

Although the data increase is different on different base models, it can be seen that the active learning strategy works regardless of the base model, and the multi-criteria sampling strategy is able to achieve better performance than the best sampling strategy, which verifies that the active learning sampling strategy proposed in this paper has universal. Through three sets of experiments, the effects of multiple active learning algorithms on different datasets with different base models are obtained, and their effectiveness and robustness are verified to achieve the expected design goals of this paper. And it is verified that the proposed multi-criteria sampling strategy outperforms the best base sampling strategy that further improving the performance of the base model and reducing the labeling cost. However, at the same time, there is also the problem that the model becomes more complex, the computational overhead increases, and it is possible to obtain only suboptimal solutions.

V. SUMMARY AND PROSPECT

As an important part of knowledge mapping, the main task of entity relation extraction is to extract the entity relation hidden in sentences. Among several popular supervised relation extraction models, the relation extraction based on pre-trained Bert language model has the best effect. This project is oriented to Chinese text, using R-BERT-based relationship extraction model as the basis, model, research based on active learning relationship extraction strategy, using a small number of tagging language materials to achieve large-scale, high-precision tagging language, material effect.

In this paper, we design a relation extraction method based on active learning, through a large number of comparative experiments, verify that it has good sample

extraction performance on different, data sets, different learning models. Specifically, aiming at the task of relation extraction, we design three different extraction strategies based on uncertainty, diversity and representativeness. Five 10,000 training sets were randomly sampled from the Baidu Brain Information Extraction Chinese dataset, including 4,000 labeled, data, each time, 200 were sampled according to different active learning strategies and handed to human experts for labeling. Then, by comparing queries, the model performance after 20 iterations is obtained-the uncertainty-based extraction strategy is the most effective, compared with random sampling, improve 6.31% of the F 1 score.

In order to improve the performance of extraction, the original three strategies are comprehensively utilized by multiple integration methods, by assigning a 2:1:1 weight to the three sampling strategies based on uncertainty, diversity and representativeness, the performance of the original optimal model is improved by 7.68%.

In this paper, an active learning strategy for relation extraction task is designed, which has been proved to be effective in relation extraction, but still has some shortcomings:

(1). In the experiment of this paper, the source of the data set is relatively single, considering the reason that the current opensource Chinese relation extraction data set is relatively few, large amounts of unlabeled data can be obtained cheaply by using third-party or self-implementing crawler systems.

(2). For active learning, the selection of initial labeled samples directly affects the training time and extraction performance of the model. In this project, the initial samples are randomly sampled, and the quality of the samples is not paid attention to, which will lead to the unreasonable distribution of the samples, which will affect the efficiency of the active learning algorithm and the final classification accuracy, the selection method of the initial samples can be improved.

(3). In constructing entity relation extraction model based on active learning, we can also design better selection strategy. For example, consider the particularity of text compared to other data formats, and design active learning strategies for semantics, or use meta-learning, to train a sampling model through learning to select effective samples.

REFERENCES

[1] C. N. Dos Santos, B. Xiang, and B. Zhou, 'Classifying relations by ranking with Convolutional neural networks', in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015. doi: 10.3115/v1/p15-1061.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.

[3] A. Sun and R. Grishman, 'Active learning for relation type extension with local and global data views', in *ACM International Conference Proceeding Series*, 2012. doi: 10.1145/2396761.2398409.

[4] G. Angeli, J. Tibshirani, J. Y. Wu, and C. D. Manning, 'Combining distant and partial supervision for relation extraction', in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014. doi: 10.3115/v1/d14-1164.

[5] T. H. Nguyen and R. Grishman, 'Relation extraction: Perspective from convolutional neural networks', in *1st Workshop on Vector Space Modeling for Natural Language Processing, VS 2015 at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015*, 2015. doi: 10.3115/v1/w15-1506.

[6] J. Wu *et al.*, 'Multi-Label Active Learning Algorithms for Image Classification: Overview and Future Promise', *ACM Computing Surveys*, vol. 53, no. 2. 2020. doi: 10.1145/3379504.

[7] Y. Wu, Y. Chen, Y. Qin, R. Tang, and Q. Zheng, 'A recollect-tuning method for entity and relation extraction', *Expert Syst Appl*, vol. 245, 2024, doi: 10.1016/j.eswa.2023.123000.

[8] T. Wu, X. You, X. Xian, X. Pu, S. Qiao, and C. Wang, 'Towards deep understanding of graph convolutional networks for relation extraction', *Data Knowl Eng*, vol. 149, 2024, doi: 10.1016/j.datak.2023.102265.

[9] A. Jose, J. P. A. de Mendonça, E. Devijver, N. Jakse, V. Monbet, and R. Poloni, 'Regression tree-based active learning', *Data Min Knowl Discov*, vol. 38, no. 2, 2024, doi: 10.1007/s10618-023-00951-7.

[10] R. Cai, X. Zhang, and H. Wang, 'Bidirectional recurrent convolutional neural network for relation classification', in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016. doi: 10.18653/v1/p16-1072.

[11] J. H. Caufield *et al.*, 'Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning', *Bioinformatics*, vol. 40, no. 3, 2024, doi: 10.1093/bioinformatics/btae104.

[12] P. Zhou *et al.*, 'Attention-based bidirectional long short-term memory networks for relation classification', in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 2016. doi: 10.18653/v1/p16-2034.

[13] L. Z. Huo and P. Tang, 'A batch-mode active learning algorithm using region-partitioning diversity for SVM classifier', *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 7, no. 4, 2014, doi: 10.1109/JSTARS.2014.2302332.

[14] H. Tang, D. Zhu, W. Tang, S. Wang, Y. Wang, and L. Wang, 'Research on joint model relation extraction method based on entity mapping', *PLoS One*, vol. 19, no. 2 February, 2024, doi: 10.1371/journal.pone.0298974.

[15] I. Lourentzou, D. Gruhl, and S. Welch, 'Exploring the Efficiency of Batch Active Learning for Human-in-the-Loop Relation Extraction', in *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 2018. doi: 10.1145/3184558.3191546.

[16] S. Wu and Y. He, 'Enriching pre-trained language model with entity information for relation classification', in *International Conference on Information and Knowledge Management, Proceedings*, 2019. doi: 10.1145/3357384.3358119.

[17] R. C. Bunescu and R. J. Mooney, 'A shortest path dependency kernel for relation extraction', in *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2005. doi: 10.3115/1220575.1220666.

[18] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Z.-P. of COLING, and undefined 2014, 'Relation classification via convolutional deep neural network', *aclanthology.orgD Zeng, K Liu, S Lai, G Zhou, J ZhaoProceedings of COLING 2014, the 25th international conference on, 2014•aclanthology.org*, Accessed: May 10, 2024. [Online]. Available: https://aclanthology.org/C14-1220.pdf