Convolution Neural Network with Active Learning for Information Extraction of Enterprise Announcements

Lei Fu[1,2], Zhaoxia Yin[1], Yi Liu[2], and Jun Zhang[3](✉)

[1] Key Laboratory of Intelligent Computing and Signal Processing,
Ministry of Education, Anhui University,
Hefei 230601, People's Republic of China
[2] PKU Shenzhen Institute, Shenzhen, China
[3] Shenzhen Securities Information, Co., Ltd., Shenzhen, China
zhangjun@cninfo.com.cn

Abstract. We propose using convolution neural network (CNN) with active learning for information extraction of enterprise announcements. The training process of supervised deep learning model usually requires a large amount of training data with high-quality reference samples. Human production of such samples is tedious, and since inter-labeler agreement is low, very unreliable. Active learning helps assuage this problem by automatically selecting a small amount of unlabeled samples for humans to hand correct. Active learning chooses a selective set of samples to be labeled. Then the CNN is trained on the labeled data iteratively, until the expected experimental effect is achieved. We propose three sample selection methods based on certainty criterion. We also establish an enterprise announcements dataset for experiments, which contains 10410 samples totally. Our experiment results show that the amount of labeled data needed for a given extraction accuracy can be reduced by more than 45.79% compared to that without active learning.

Keywords: Text classification · Active learning
Convolutional neural networks · Enterprise announcements

1 Introduction

At present, information extraction has become an important branch of the NLP field. The task of information extraction is to obtain target information accurately and quickly from a large amount of data and improve the utilization of information. The information extractions of enterprise announcements help the users identify concerns quickly. Therefore, this paper addresses information extraction task for enterprise announcements which is a type of document that publicly informs the society of important issues and extracts the information about investment and other specifically. We extract the key information through text classification. There are many methods for text classification. Currently, the mainstream method implements text classification through deep learning.

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 330–339, 2018.
https://doi.org/10.1007/978-3-319-99501-4_29

There is no doubt that deep learning has ushered in amazing technological advances on natural language processing(NLP) researches, applied to various tasks such as text classification [1], machine translation [2], document summarization [3]. But in the deep learning environment, there is a bottleneck that is manual collection of sample labels is expensive and time-consuming. Convolutional neural networks (CNN) [4] is common model in deep learning and many natural language processing tasks and has great classification performance. In order to obtain great classification performance, a large amount of manually-labeled data is needed. Therefore, the low efficient manual labeling work has become a problem that restricts the development of text classification tasks. Active learning [5, 6] is motivated for solving the problem. Active learning is an iterative process of selecting useful samples as training data. And the size of the training data set should be as small as possible while maintaining classification performance. Therefore, the kernel of active learning is to choose a useful sample set. This paper proposes an active learning method to improve the effect of deep learning for text classification task.

Based on the deep learning architecture, this paper proposes a novel active learning algorithm, which is to judge the useful data according to the category probability strategy. It solves the problem that the applying deep learning to text classification require a lot of manual data annotation.

In this paper, Sect. 2 describes the related work of this study. Section 3 gives a description of the principles of CNN and active learning methods. Section 4 introduces the experimental results and analysis. Section 5 concludes the paper and outlines the future work.

2 Related Work

At present, there are a lot of research about active learning. Most of the early work can be found in the classical survey of Settles [7]. It covers acquisition functions such as information theoretical methods [8]. And Bayesian active learning method typically uses a non-parametric model like Gaussian process to estimate the expected improvement by each query [9] or the expected error after a set of queries [10]. [9] presented a discriminative probabilistic framework based on Gaussian Process priors and the Pyramid Match Kernel and introduced an active learning method for visual category recognition based on the uncertainty estimates provided by the GP-PMK. [10] presents an active learning method that directly optimizes expected future error. A recent approach by Gal & Ghahramani [11] shows an equivalence between dropout and approximate Bayesian inference enabling the application of Bayesian methods to deep learning. The Support Vector Machines [12] of active learning method presented two novel multi-label active learning strategies, a max-margin prediction uncertainty strategy and a label cardinality inconsistency strategy, and then integrate them into an adaptive framework of multi-label active learning. The Query by committee [13] active learning method is to train a committee of learners and query the labels of input points where the committee's predictions differ, thus minimizing the variance of the learner by training on input points where variance is largest. The Expectation-Maximization (EM) [14] with active learning (uses a modified QBC) for text classification modify the

Query-by-Committee (QBC) method of active learning to use the unlabeled pool for explicitly estimating document density when selecting examples for labeling. Then active learning is combined with Expectation-Maximization in order to "fill in" the class labels of those documents that remain unlabeled.

In our research work, the active learning is mainly applied to the image [15] and rarely applied to the text for the deep learning method. Proposed in this paper, the certainty criterion applies active learning to text classification tasks in deep learn fields.

3 Proposed Method

Figure 1 shows the framework diagram of CNN based on active learning. This section mainly introduces the classification method of CNN and the principle of active learning method. Convolution neural network can reduce the parameter quantity through the local connection and weight sharing to greatly reduce training complexity and overfitting. Meanwhile weight sharing also gives the convolutional network tolerance for translation. Active learning methods can effectively reduce the number of training samples, therefore significantly reducing training time and manual tagging workload.

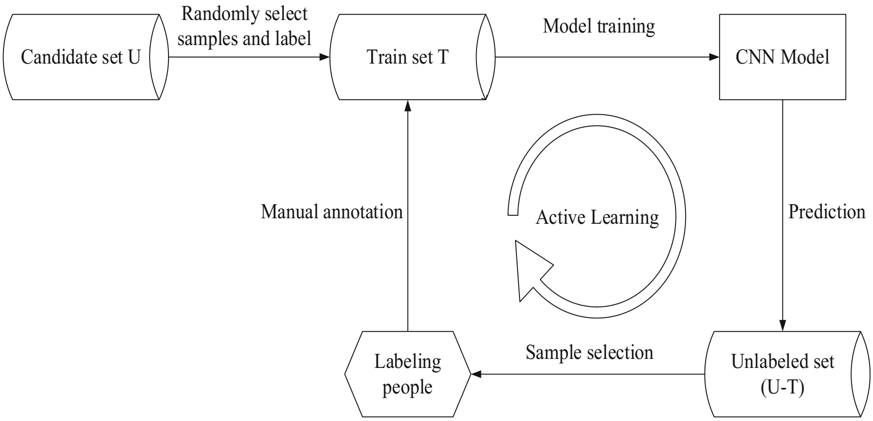

Fig. 1. CNN based on active learning

3.1 Convolutional Neural Network

CNN is a deep neural network and has made a major breakthrough in computer vision and speech recognition, and mainly contains the convolution layer and the pool layer. By convolutional operations at different scales are implemented on long text, more comprehensive features can be extracted.

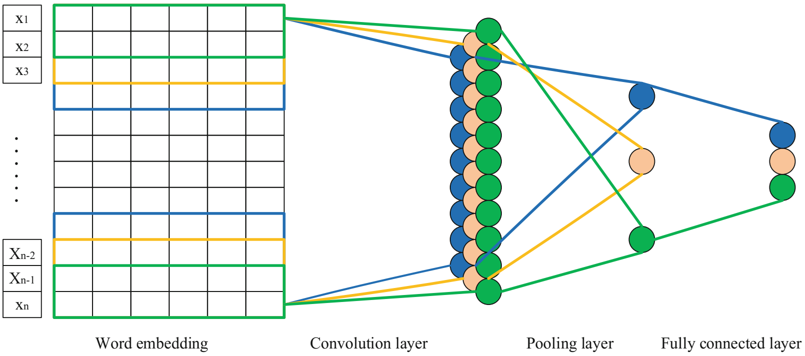

Fig. 2. CNN model diagram

Figure 2 shows the framework of the CNN. x_i indicates the k-dimensional word embedding of the *i-th* word of a text. A text $x_{1:n}$ which is length n is represented as:

$$x_{1:n} = x_1 \oplus x_2 \oplus \cdots \oplus x_n \tag{1}$$

$\oplus$ indicates a concatenation operation. In general, let $x_{i:i+j}$ indicate $x_i, x_{i+1}, \cdots, x_{i+j}$. A filter is a window of size h words and produces a new feature by a convolution operation, a convolution operation corresponding to a filter $w \in R^{hk}$. For example, feature c_i is generated from a window size h words $w_{i:i+k-1}$ by:

$$c_i = f(w \cdot x_{i:i+k-1} + b) \tag{2}$$

Here, b denotes the bias term, f denotes a nonlinear function such as rectifier or tanh. Applying a filter to each window $\{x_{1:h}, x_{2:h+1}, \cdots, x_{n-h+1:n}\}$ in the text to generate a set of feature maps:

$$c = [c_1, c_2, \cdots, c_{n-k+1}] \tag{3}$$

$c \in R^{n-h+1}$. Then the largest feature map of the group of feature maps c is selected to represent the group of features by the maximum pooling operation.

A filter extracts a feature. Therefore, multiple features are extracted by multiple filters (windows of different sizes). These features are connected in the full connection layer to get the text feature vector and sent it to Softmax layer to get its corresponding category. The model uses a cross-entropy function as a loss function, which measures the probability error of each independent classification task.

3.2 Active Learning

At present, the supervised algorithms require a lot of labeled training data, and the result of the experiment is affected the quality of the data in deep learning. This paper introduces active learning into the CNN to classify long text. In essence, active learning is an iterative training process of selecting a useful sample from unlabeled sample data and obtaining its label.

The three sample selection methods based on certainty criterion proposed in this paper chooses the next sample to be annotated from the unmarked sample ***U*** based on certainty criterion, which has three kinds of sample selection methods.

The First Method: Selecting the sample with probability range of $[1/C - a, 1/C + b]$. The probability of the sample belongs to this range is the uncertainty sample. And if the sample is wrongly predicted, it has a large impact on the classifier and meets the selection requirements.

The Second Method: Selecting the sample with probability range of $([0, c] \cup [d,1])$. The probability of the sample belongs to this range is the certainty sample. And if the sample is wrongly predicted, it has a large impact on the classifier and meets the selection requirements.

The Third Method: this method is to select $\propto$ of the first method and $(1\text{-}\propto)$ of the second method, where $\propto$ is obtained based on engineering experience.

The total number of categories is C, and the constants of a, b, c, and d are based on engineering experience. Using these three probabilistic selection methods, each sample selected is a useful sample data, so it also has the greatest influence on the classifier. The influence of the remaining samples on the classifier is gradually weakened.

The specific steps of active learning based on certainty criterion are as follows:

```
Probability Selection Strategy
    Input: Uncategorized the candidate sample set $U$.
    Output: Classifier $fc$.
    Begin:
    Step 1. Selecting $i$ samples from the candidate sample set $U$ and correctly
            labelling their categories to construct the initial training set $T$;
    Step 2. Using the training set $T$ to train the classifier $fc$;
    Step 3. Using the classifier $fc$ to classify the sample ($U$-$T$) remaining in the
            candidate set, and selecting the sample that meets the requirements
            (Within a specified range of probabilities and label errors predicted)
            for annotation;
    Step 4. Using labeled data and training set $T$ to train a new classifier $fc$;
    Step 5. If the accuracy reaches $\beta$, the algorithm terminates and returns $fc$;
            Otherwise returns to step 2);
    End.
```

For active learning, the classifier doesn't passively accept the data provided by the user, but instead actively asks the user to label the sample that is selected by the current classifier. Through continuous selection of indistinct samples to iteratively train obtain a satisfactory classifier. Theoretically, active learning can significantly reduce the number of required samples compared to random selection with similar experimental results [16].

4 Experiment

For verifying the validity of our method, we conduct a series of experiments on our self-built enterprise announcements dataset. Note that we apply the CNN as the basic deep learning classification method.

4.1 Data Set

We crawl the enterprise announcements as our dataset on the Internet. And each enterprise announcement is segmented to form the corresponding text segment according to the title. The detail is listed in Table 1. In order to the convenience of experiment, the text segments are annotated manually in advance. In other words, the categories of the text segments are divided into Investment and Other labels. The corresponding meaning of the labels is listed in Table 2.

Table 1. Data set

Date set	Total (number)	Investment (number)	Other (number)
Train	8554	2669	5885
Valid	916	101	815
Test	940	120	820

Table 2. Corresponding meaning of the labels

Label	Corresponding meaning
Investment	The content of the text block is related to investment
Other	Other represents text blocks that have nothing to do with investment

4.2 Experimental Setting

The dimension of the CNN model is *Dim* = 64, the number of convolution kernel is 3, and the convolution kernel size are 2, 3 and 4. According to the observation of all enterprise announcements, their length is about 400 words. Therefore, all enterprise announcements are fixed to 400 words. If the length of sample is greater than 400

Table 3. Method and probability values

	Probability one	Probability two	Probability third
The first method	(0.1,0.9)	(0.125,0.875)	(0.15,0.85)
The second method	[0,0.1] ∪ [0.9,1]	[0,0.1] ∪ [0.95,1]	[0,0.05] ∪ [0.9,1]
The third method	(0.125,0.875) ∪ ([0,0.1] ∪ [0.95,1])	(0.125,0.875) ∪ ([0.0.05] ∪ [0.9,1])	(0.15,0.85) ∪ ([0,0.1] ∪ [0.95,1])

words, we select the foregoing 400 words. Otherwise, tag <pad> is added for complementation until achieving 400 words. Besides, for our dataset, this paper proposes three group values for each method in Table 3.

Through our empirical study, the parameters are set the appropriate value. The α of the third method is 0.5. When the learning rate of *lr* is set 0.01, the effect of model is best. Furthermore, the F1-measure don't increase significantly after 30 consecutive epochs and the model stopped updating iterations. In this experiment, Word2vec of Google Company is used to pre-train the unlabeled 700 MB Chinese Sogou corpus into a word embedding table of $Dim = 64$. Therefore, the trained words embedding can represent the words accurately.

4.3 Evaluation Index

General text classification evaluation criteria use accuracy (P), recall (R) and F1-measure (F) as an indicator. Accuracy rate is the correct rate of information related to the need to pay attention; recall rate is the proportion of information retrieved; the F1-measure can be thought of as a weighted average of model accuracy and recall, with a maximum of 1 and a minimum of 0. F1-measure is calculated as follows:

$$F_1 = \frac{2TP}{2TP + FN + FP} \tag{4}$$

In this experiment, TP indicates that the predicted sample label of model is Investment, and the actual label is Investment; FP indicates that the predicted sample label of model is Investment, and the actual label is other; FN indicates that the predicted sample label of model is other, and the actual label is Investment; TN indicates that the predicted sample label of model is other, and the actual label is other. In this paper, the F1-measure of experimental result is represented.

4.4 Experimental Results and Analysis

In the Tables 4, 5 and 6, the first row is the probability value and the first column is the number of iterations, the middle content is the F1 value (a used percentage of all data). x indicates the prediction probability of sample.

Table 4. Experimental results of first method

	0.1 < x < 0.9	0.125 < x < 0.875	0.15 < x < 0.85
Step 1	0.79 (42.66)	0.79 (42.66)	0.79 (42.66)
Step 2	0.77 (46.20)	0.78 (45.24)	0.78 (45.28)
Step 3	0.79 (51.43)	0.80 (47.99)	0.80 (50.15)
Step 4	0.80 (53.59)	0.82 (50.22)	0.81 (56.09)
Step 5	0.81 (56.04)	0.81 (52.26)	0.84 (58.02)
Step 6	0.83 (58.23)	0.83 (54.58)	**0.85 (59.40)**
Step 7	**0.85 (58.67)**	**0.85 (55.62)**	0.84 (60.97)
Step 8	0.84 (60.10)	0.85 (57.21)	0.84 (62.68)

Table 5. Experimental results of second method

	x <= 0.1 or x >= 0.95	x <= 0.05 or x >= 0.9	X <= 0.1 or x >= 0.9
Step 1	0.79 (42.66)	0.79 (42.66)	0.79 (42.66)
Step 2	0.78 (45.92)	0.77 (45.45)	0.77 (46.08)
Step 3	0.83 (47.99)	0.82 (48.01)	0.81 (48.19)
Step 4	**0.85 (55.70)**	**0.85 (56.18)**	**0.85 (55.94)**

Table 6. Experimental results of third method

	(x <= 0.1 or x >= 0.95) or 0.125 < x < 0.875	(x <= 0.05 or x >= 0.9) or 0.125 < x < 0.875	(x <= 0.1 or x >= 0.95) or 0.1 < x < 0.9
Step 1	0.79 (42.66)	0.79 (42.66)	0.79 (42.66)
Step 2	0.79 (44.54)	0.77 (44.41)	0.78 (45.56)
Step 3	0.80 (48.23)	0.80 (49.32)	0.81 (51.46)
Step 4	0.82 (50.41)	0.81 (51.18)	0.83 (58.23)
Step 5	0.83 (52.84)	0.83 (52.91)	0.84 (59.11)
Step 6	0.84 (54.72)	**0.85 (54.21)**	**0.85 (59.82)**
Step 7	**0.85 (55.41)**	0.83 (56.14)	0.84 (60.02)

When using all of the training data, the F1 value of results 0.85, which is compared with our methods as the baseline method. Tables 4, 5 and 6 are the experimental results of our three methods.

Table 4 shows the experimental results of our first method by using three different sets of probability range values. From Table 4, we can find that only 55.62% of the sample data can arrive the expected experimental result 0.85, when the selection probability range is (0.125, 0.875). The effect of this experiment is the best for the first method.

Table 5 shows the experimental results of our second method. And we also apply three different sets of probability range values. From Table 5, we can see that only 55.70% of the sample data can arrive the expected experimental result 0.85, when the selection probability range is [0, 0.1] ∪ [0.95, 1]. The effect of this experiment is the best for the second method.

Table 6 is the experimental results of our third method, which select 50% of the first method and 50% of the second method. From Table 6, it can be observed that only 54.21% of the sample data can arrive the expected experimental result 0.85, when the selection probability range is ([0, 0.05] ∪ [0.9, 1]) ∪ (0.125, 0.875)). The effect of this experiment is the best result of third method and also the best result of these three methods. The result of experiment shows that our active learning approach requires 45.49% less training data. When F1 reaches a certain value, F1 tends to a stable state. In the meantime, there are few samples taken through the category probability strategy each time. Furthermore, those experiment results indicate we can achieve the expected experimental results without using all the data. As these useful samples, selected for each time, have a large impact on the classifier, we can use a small amount of data to achieve the same effect as a large amount of data.

5 Conclusion

In this paper, we propose a certainty criterion under the framework of deep learning to solve the problem of requiring a lot of manual data annotation. The certainty criterion has three selection method based on certainty criterion of classification. Through these three selection methods, the convolutional neural network is trained on the selected data iteratively, so that the model can achieve the expected results. In the experiment, it can be found that 45.49% of the manually labeled work can be saved in the best result. In our further work, we will investigate more text classification algorithms combined with our proposed method and apply our proposed method to other areas.

Acknowledgments. This work is partially supported by Shenzhen Science & Research projects (No: JCYJ20160331104524983) and Key Technologies Research & Development Program of Shenzhen (No: JSGG20160229121006579). We thank the reviewers for the insightful comments.

References

1. Yogatama, D., et al.: Generative and discriminative text classification with recurrent neural networks (2017). arXiv preprint, arXiv:1703.01898
2. Tu, Z., et al.: Modeling coverage for neural machine translation (2016). arXiv preprint, arXiv:1601.04811
3. Cao, Z., et al.: Improving multi-document summarization via text classification. In: AAAI (2017)
4. Conneau, A., et al.: Very deep convolutional networks for natural language processing (2016). arXiv preprint, arXiv:1606.01781
5. Joshi, A.J. Porikli, A.J., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2372–2379 (2009)
6. Tur, G., Hakkani, D.: Combining active and semi-supervised learning for spoken language understanding. Speech Commun. **45**(2), 171–186 (2005)
7. Settles, B.: Active learning literature survey. Univ. Wisconsin, Madison **52**(55–66), 11 (2010)
8. MacKay, David J.C.: Information-based objective functions for active data selection. Neural Comput. **4**(4), 590–604 (1992)
9. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with Gaussian processes for object categorization. In: ICCV (2007)
10. Roy, N., McCallum, A.: Toward optimal active learning through monte carlo estimation of error reduction. In: ICML (2001)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning (2016)
12. Li, X., Guo, Y.: Active learning with multi-label SVM classification. In: IJCAI (2013)
13. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM (1992)

14. McCallumzy, A.K., Nigamy, K.: Employing EM and pool-based active learning for text classification. In: Proceedings of the International Conference on Machine Learning (ICML), pp.359–367 (1998). Citeseer
15. Wang, K., et al.: Cost-effective active learning for deep image classification. IEEE Trans. Circ. Syst. Video Technol. **27**(12), 2591–2600 (2017)
16. Dasgupta, S.: Coarse sample complexity bounds for active learning. In: Advances in Neural Information Processing Systems (2006)