An Information Extraction Approach Based on Domain Language Model

Jing Zhu*, Yongcui Deng, Yiwen Zhou, Dewang Sun and Ruibin Mao

Shenzhen Securities Information Co., Ltd, Shenzhen 518022, China
*zhujing@cninfo.com.cn

Abstract. The information extraction (IE) of listed company announcements is the basis for structuring company announcement information. It can improve the efficiency of investors and regulators in tracking and regulating the market. We introduce BERT for listed company announcements text mining tasks called caBERT, based on BERT architecture, we pre-train caBERT with listed company announcements corpus on the performance of information extraction tasks and achieve competitive results, and in personnel-change-related announcement our best submission achieves the F1 score of 97.27%.

Keywords: Listed Company Announcements, BERT, Information Extraction.

1 Introduction

With the development of Fintech and the continuous expansion of the global capital market, in the financial field, there are huge amounts of data generated every day, which is in sharp contrast to the limited human resources and the limit ability of human brain to process information. Therefore, relying only on traditional manual methods has been unable to cope with the needs of investment analysis, risk control, financial supervision and event correlation. It is urgent to introduce new technologies to improve the efficiency of information processing. Emerging technologies such as big data, natural language processing, knowledge graph have been actively used in financial analysis and financial regulation.

Information extraction is the process of identifying a particular class of events or relationships and extracting related parameters of these events and relationships in a natural language text [1,2]. Driving by evaluation conference such as Message Understanding Conference (MUC), Automatic Content Extraction (ACE) and Text Analysis Conference (TAC), the research on information extraction technology has been flourishing. Promoted by the Chinese Information Society, the China Conference on Knowledge Graph and Semantic Computing (CCKS) launched the Chinese information extraction evaluation since 2016. By 2019, it launched the evaluation for information extraction of listed company announcement for the first time. These actions have strongly promoted the development of Chinese information extraction technology.

Listed companies announcements are mainly formatted in PDF. PDFs are stored by characters and locations, there is no clear distinction of paragraphs or tables, thus add difficulties in analyzing them. Besides, there are various categories of listed company announcement, most of them have a chapter hierarchy. However, the characteristics of mixing table and text and semantic nesting of the announcements pose challenges to information extraction technology. According to these characteristics, we purposed the announcement information extraction system for listed companies which is based on the multi-layer information extraction framework at the chapter level and integrates various information extraction techniques to realize automation

2 Related work

At present, there are many related studies on information extraction methods, and many of them focus on three aspects : traditional method, sequence labeling and dependency tree.

Traditional information extraction mainly relies on manual definition of relevant features, which includes the method based on expert knowledge [6], rules [7,8] and dictionaries [9]. The early IE system is a typical example of the use of expert knowledge, and it creates a language knowledge in the form of rules or patterns to detect and extract target information from text. ATRANS [7] and JASPER [8] are classic methods based on rule. ATRANS[7]is a method based on bank knowledge and script framework which could extract information from information about interbank capital transfers, while JASPER[8] is a method based on template-driven approach, and it could extract the information of corporate earnings report from relevant text; UMASS/MUC-4[9] is a system based on sentence analysis and the form of least domain lexicon, which uses dictionary method for functional implementation.

The sequence labeling method is a method to transform information extraction into solving sequence labeling problem, its original solution was graph probability model, including HMM [10], MEMM [11], CRF [12] , Semi-CRF [13], etc. With the popularization of deep learning methods and the excellent contextual modeling ability of recurrent neural networks, many scholars have achieved good performance in solving sequence labeling problem, like RNN+CRF[14], LSTM+CRF[15], Bi-LSTM+CRF[16,17], and ENCODER-LABLER[18]. Sequence labeling method is still one of the most commonly used methods for information extraction so far.

The dependency tree method introduces syntactic features coding in extraction, which can effectively improve the extraction accuracy. Fader [19] introduced two simple syntactic and lexical constraints to represent the binary relationship expressed by verbs, which solved the problem of information loss and inconsistency. Miwa [20] realized end-to-end extraction by obtaining the structural characteristics of the word sequence and dependency tree. Peng [21] proposed to obtain the syntactic expression

of words by graph-LSTMs, which combined the features of word vectors, and finally extracted the multivariate relations in paragraphs by classifier.

However, these methods only take the syntactic features as one of the feature inputs of the extraction task, and don't directly apply the semantic tree structure for training. Essentially, they are the extension of feature engineering and cannot obtain the semantic structure well.

3 IE System of Listed Company Announcements

3.1 System Framework

In order to achieve accurate parsing, we design an extraction method using pipeline mode, which is divided into three steps: PDF parsing, structure identification and information extraction. Since the announcement is the semi-structured text, the text contains a large number of directory structure tags, which divide the announcement into several chapters, and each chapter contains several smaller chapters. Therefore, information is located from the chapters to effectively reduce errors. The system processing flow is showing in Fig.1.

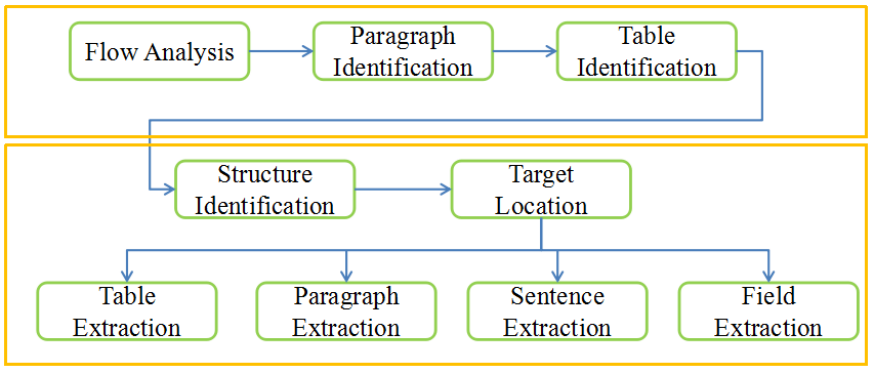

Fig. 1. System processing flow. Flow analysis, paragraph identification and table identification mainly divide the PDF announcement into texts and tables, then identify catalogue. Next, preliminary locating the extracting targets according to the title and text features. Finally, performing the ultimate extraction task.

3.2 PDF Parsing

Current PDF information extraction is to use tools to convert PDF files into other easy-to-handle formats, such as HTML, XML, and Word. However, such format conversion may result in information loss. We directly parse the PDF source file to get the raw information of the text and tables, including character, font, font size, font position and other information. And then we merge the characters into a text block.

Finally, we further merge the processed text blocks into paragraphs based on the feature of the paragraphs.

3.3 Domain Language Model

BERT[1] is an abbreviation for Bidirectional Encoder Representation from Transformers, its architecture is a multi-layer bidirectional Transformer encoder. It uses "masked language model" and "next sentence prediction" tasks to learn word and sentence representations respectively. Unlike GPT[2] and ELMo[3], BERT[1] is pre-trained bidirectional representations from a large amount of unlabeled corpus and joint results both from left and right context on each layer. As a result, additional another output layer on BERT and fine-tuning can create state-of-the-art models in many NLP tasks.

The google Chinese BERT-base pre-trained model is based on Chinese wiki corpus, which contains the basic content of topics such as biography, history, geography, society, culture, science, technology, food and mathematics, also involves some popular characters and events. When it applys to the listed company announcements, we should consider the language expression characteristics of individual events in the securities field. In this respect, the announcement language shows different characteristics from wikis and social news. In addition, the description of the elements in the securities field generally consists of time, subject and values, which are different from those of common sense. Therefore, for the natural language processing of listed company announcements, it is necessary to train the language model of the securities field. Based on the corpus of the listed company announcement, we train the domain language model caBERT. We will fine-tune and compare the task based on the caBERT below.

3.4 Field Extraction

Based on the domain language model, the extracted information can be identified. However, there are some complex sentences in the text of the announcement, such as partial indexing, multi-finger, and nested structure. Since the information obtained by sequence labeling are discrete and have no corresponding relationship, we use syntactic tree combination rules or combine hierarchical information according to the characteristics of the announcement. In this way, the information is processed into a triplet form like (object, key, value), which can achieve the corresponding result of the field extraction result.

4 Experiment

4.1 Data Preprocessing

The CCKS 2019 task5- extract information from personnel-change-related announcements provided 617 announcements as training dataset. According to the given

PDF files, we converted them to text file and extracted related sentences from announcements. Then according to the training data, the corpus is re-tagged to obtain the pre-labeled data set, for each character in the match, the entities were encoded in BIO tagging scheme. For the existence of multiple and missing labels in the re-tagging process, we create rules to identify mistakes and finally obtain the training set.

This paper adopts precision (P), recall (R) and F1 score (F1) as its evaluation indicator, its calculation method as follow:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2PR}{P + R}.$$

For a specific entity, TP represents the number of correct identifications of the specific entity, FP represents the number of misjudgments of identifying other entity as the specific entity, and FN represents the number of misjudgments of identifying the specific entity as other entity.

4.2 Pretrain Language Model

We pre-trained a domain language model for tasks in the securities. The corpus contains 598 varieties listed company announcements in the past three years. Considering the computational complexity and expensive of BERT-large, we only used the parameters of BERT-base for training and we trained eight days with eight NVIDIA V100 (32GB) GPUs. The fine-tuning was applied to personnel-change-related information extraction tasks. Figure 1 is showing the structure of BERT which applied to NER tasks.

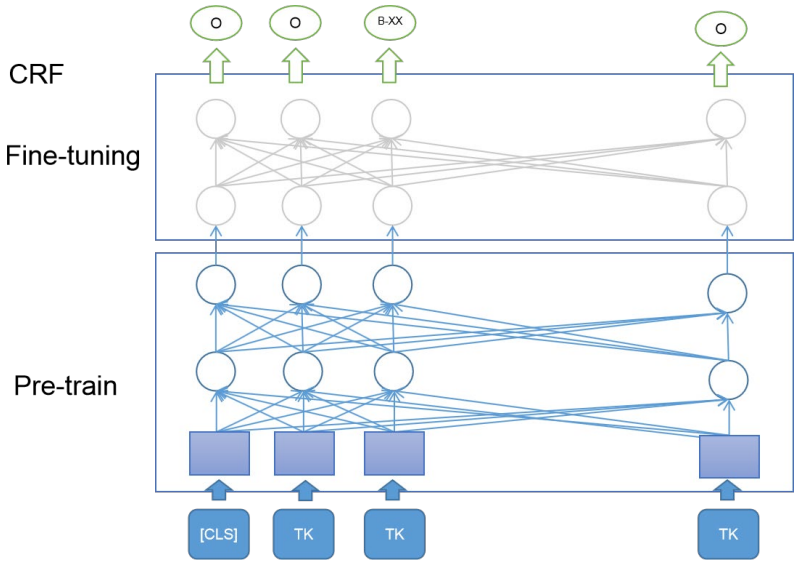

Fig. 2. BERT-CRF structure.

4.3 Analysis of Experiment

BiLSTM-CRF[16,17] is the most popular and outstanding performance sequence labeling model in recent years. BiLSTM-CRF was applied first, the parameters were set as follows: batch-size = 128, maximum sentence length =120, and Word2vec dimension= 150. However, the final average F1 score on test set was 80.49%, which suggests the adaptability and practicability are not enough. In order to improve the accuracy of entity recognition, BERT-CRF was applied with google Chinese BERT-base pre-trained language model and the result showed that BERT-CRF model achieved better result on the same training set, the average F1 score of BERT-CRF model was 13.14% higher than BiLSTM-CRF.

Although the result of BERT-CRF model has been significantly improved, the promotion space in the recognition effect of the reason for resignation and the position of resignation still remains. According to the analysis of result from BERT-CRF model, it was found that the mistaken identification of reason for resignation and position of resignation were mainly due to the length of entity text(composed of more than 5 words). Thus, the model only predicted part of words in the whole entity. After statistics on the training set, reason for resignation and post of resignation only took a small percentage of the whole corpus. Therefore, the corpus of the long entity text is especially enhanced as follows:

1. Extract the entities with more than 5 words in the training set and form a list;
2. Extract sentences contain the entities to be replaced, and copy each sentence into 10 copies to form the sentence set;
3. For each sentence in step 2, randomly select the entity in step 1 to replace;
4. The enhanced training set is obtained by combining the original training set with the newly enhanced training set.

Table 1. Comparison before and after corpus enhancement.

	Training set	Development set	Test set
Number of sentence before corpus enhancement	2351	295	293
Number of sentence after corpus enhancement	5901	295	293

The length and count number of reasons for resignation was statisticsed before and after data enhancement, as well as the position of resignation, and the corpus imbalance problem was effectively solved.

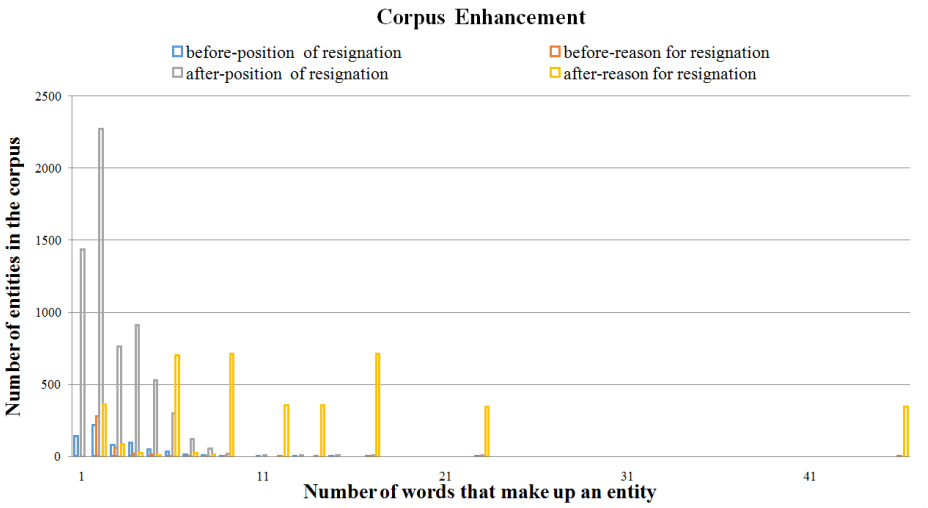

Fig. 2. Corpus enhancement.

After corpus enhancement, the model performance has been significantly improved on reason for resignation and the position of resignation. Then, we used BERT-CRF + caBERT and fine-tune it. The results showed that the domain language model can effectively improve the efficiency of recognition in certain professional field. The different results were shown below.

Table 2. F1 score on test set. POR, NOR, GOR, RFR, POS, NOS and GOS represent position of resignation, the name of resignation, the gender of resignation, the reason for resignation, the position of successor, the name of successor, the gender of successor, respectively.

	Average	POR	NOR	GOR	RFR	POS	NOS	GOS
Bilstm-CRF	80.49	85.56	72.53	75.27	**95.83**	83.87	61.11	62.86
Bert-CRF	93.63	93.40	98.00	98.00	82.57	97.14	97.96	97.96
Bert-CRF +enhance	96.78	97.44	96.08	96.08	94.00	**100.00**	**100.00**	**100.00**
Bert-CRF +enhance +caBERT	**97.27**	**97.56**	**96.58**	**97.01**	95.82	98.14	**100.00**	**100.00**

5 Conclusion

This paper introduces a information extraction system, which can extract listed company announcement. The domain language model in this paper has a good performance in entity recognition. In the CCKS 2019 task5- extract information from per-

sonnel-change-related announcements, we achieved F1 score of 95.78% which ranked the first. We will focus on more information extraction technologies in our future work.

References

1. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805, (2018).
2. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, OpenAI. (2018).
3. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In NAACL. (2018).
4. Zhang Jun. Commercial Information Service and its Implementation Strategies in Securities Market[D]. Wuhan: Wuhan University, (2005).
5. Ren Pengfei. Research on the Impact of Listed Company Announcement Event on Stock Price[D]. Jinan: Shandong University, (2016).
6. Piskorski J, Yangarber R. Information extraction: Past, present and future[M]. Multi-source, multilingual information extraction and summarization. Berlin, Heidelberg: Springer, 23-49 (2013).
7. Lytinen S L, Gershman A. ATRANS Automatic Processing of Money Transfer Messages[C]. In: AAAI. 1089-1093 (1986).
8. Andersen P M, Hayes P J, Huettner A K, et al. Automatic extraction of facts from press releases to generate news stories [C]. In: Proceedings of the third conference on Applied natural language processing. 170-177 (1992).
9. Lehnert W, Cardie C, Fisher D, et al. University of Massachusetts: MUC-4 test results and analysis[C]. In: Proceedings of the 4th conference on Message understanding. 151-158 (1992).
10. Leek T R. Information extraction using hidden Markov models[D]. San Diego: University of California, San Diego, (1997).
11. Mccallum A, Freitag D, Pereira F C. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]. In: Proceedings of the Seventeenth International Conference on Machine Learning, CA, USA. CA,USA: Morgan Kaufmann Publishers Inc, 591-598 (2000).
12. Lafferty J, Mccallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. In: Proceedings of the Eighteenth International Conference on Machine Learning, CA,USA. CA,USA: Morgan Kaufmann Publishers Inc, 282-289 (2001).
13. Sarawagi S, Cohen W W. Semi-markov conditional random fields for information extraction[C]. In:Advances in neural information processing systems, Vancouver,Canada. MA,USA: MIT Press Cambridge, 1185-1192 (2005).
14. Mesnil G, He X, Deng L, et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[C]. In:Interspeech 2013, Lyon, France. New York,USA: eprint arXiv, 3771-3775 (2013).
15. Yao K, Peng B, Zhang Y, et al. Spoken language understanding using long short-term memory neural networks[C]. In:2014 IEEE Spoken Language Technology Workshop (SLT), NV, USA. USA:IEEE, 189-194 (2014).

16. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, (2015).
17. Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]. In:Proceedings of NAACL 2016, California,USA. PA,USA:Association for Computational Linguistics, 260-270 (2016).
18. Kurata G, Xiang B, Zhou B, et al. Leveraging sentence-level information with encoder lstm for semantic slot filling[C]. In: EMNLP 2016, Texas,USA. New York,USA: eprint arXiv, 2077- 2083 (2016).
19. Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction[C]. In:Proceedings of the conference on empirical methods in natural language processing, Edinburgh, United Kingdom. PA,USA: Association for Computational Linguistics, 1535-1545 (2011).
20. Miwa M, Bansal M. End-to-end relation extraction using lstms on sequences and tree structures[C]. In:Association for Computational Linguistics (ACL),2016, Berlin,Germany. PA,USA: Association for Computational Linguistics, 1105-1116 (2016).
21. Peng N, Poon H, Quirk C, et al. Cross-sentence n-ary relation extraction with graph lstms[J]. Transactions of the Association for Computational Linguistics, 5,101-115 (2017).