An Event-oriented Model with Focal Loss for Financial Event Subject Extraction

Kunxun Qi, Jianfeng Du*, Jinglan Zhong, ZhenJie Chen, Hanying Lai, and Langlun Chen

School of Computer Science and Technology,
Guangdong University of Foreign Studies, Guangzhou 510006, China
*Corresponding author: jfdu@gdufs.edu.cn

Abstract. In this paper, an event-oriented neural model with focal loss is proposed to recognize the financial event subject within a text according to an event type. This model is basically enhanced from a pre-trained language model on two main aspects. On one hand, an attention mechanism is proposed to learn the interactive representation between a text and an even type. On the other hand, a three-stage fine-tuning mechanism based on focal loss is proposed to train the model. The proposed model is evaluated on a dataset about Chinese financial news from C-CKS 2019. Experimental results show that the model achieves a 92.58% F1-score on the validation set and a 83.78% F1-score on the test set.

Keywords: event extraction · named entity recognition · machine reading comprehension

1 Introduction

Event extraction is a challenging task in Nature Language Processing (NLP). It aims at discovering event mentions and extracting events which contain event triggers and event arguments from texts [29]. Financial event subject extraction is a special case of event extraction, which can provide valuable information for investment analysis and asset management. More precisely, financial event subject extraction aims to recognize event subject entities within a text according to a given event type.

There are two tasks in NLP that are highly related to financial event subject extraction, namely Named Entity Recognition (NER) [28] and Machine Reading Comprehension (MRC) [17]. NER aims to recognize the entities like person, organization and location within a text. MRC aims to extract the answer of a given question within a document or multiple documents. By treating the given event type as the question in MRC and the event subject as the answer in MRC, existing models for MRC could be adapted to financial event subject extraction.

Recently pre-trained language models like BERT [7] have become popular in tackling NLP tasks especially the MRC task. Take BERT for example, a direct adaptation of BERT to financial event subject extraction is adding an answer pointer layer [25] beyond BERT to predict the start position and the

end position of the event subject. Similar adaptations of pre-trained language models have been shown to work well for the MRC task [7]. However, such a direct adaptation of BERT has the following limitations. On one hand, from the same given text the extracted event subjects can be different according to different given event types. The direct adaptation of BERT does not capture the dependency between event subjects and event types. On the other hand, the distribution of event types is imbalanced. Like other statistical models, BERT tends to predict event subjects more accurately for event types that have more training examples. It usually has a rather poor performance for minority event types that have a small number of training examples.

To address the above two limitations, we enhance the direct adaptation of a pre-trained language model on two main aspects. On one hand, we propose an attention mechanism to learn the interactive representation between the given text and the given even type. Our proposed model with this attention mechanism is depicted in Fig. 1. In this model, trainable event type embeddings are introduced to compute an attentive representation of the given text based on the contextual representation of the given text that is generated by a lexical encoder such as BERT or other pre-trained language models. Afterwards, the interactive representation is generated by concatenating the attentive representation with the contextual representation. On the other hand, we propose a three-stage fine-tuning mechanism to make a better prediction for hard examples. The hard examples are those examples for which a learnt model is hard to make the correct predictions. They are usually examples on the minority event types. We employ a new loss function named focal loss [13] in the minimization goal function. As illustrated in Fig. 2, focal loss is a dynamically scaled cross-entropy loss, where the scaling factor decreases as the confidence on the correct prediction increases. In this way the learning course will focus more on those examples that have lower confidences on the correct predictions. Based on this observation, in the proposed fine-tuning mechanism we set different values of the hyper-parameter in focal loss for different stages, in order to make the learning course pay more and more attentions to hard examples in consecutive stages.

The above enhanced model predicts answer spans for extracting the event subject. There can be multiple entities in an event subject. To pick out multiple entities in the event subject, we develop an algorithm for postprocessing predicted answer spans. This algorithm selects up to three predicted answer spans whose prediction probabilities are not much less than the prediction probability of the top predicted answer span. Afterwards, the algorithm widens the selected answer spans according to the connection characters "、", "，", "和" and "以及" therein, and then extracts target entities from the enlarged answer spans by splitting answer spans with the connection characters.

Our proposed method is evaluated on the dataset about Chinese financial news from CCKS 2019. Experimental results show that the method achieves a 92.58% F1-score on the validation set and a 83.78% F1-score on the test set. In addition, the comparison results for ablation study further demonstrate the effectiveness of our proposed enhancements.

2 Related Work

Event extraction has gained much attention recently. Previous methods for event extraction can be grouped in three main categories, namely statistical methods, pattern-based methods and hybrid methods [10]. Statistical methods include traditional machine learning methods that rely on feature engineering [1, 11, 19] as well as neural network based methods that extract features automatically [5, 16, 15]. The pattern methods usually build a relatively complete pattern library and use a semi-automatic method to build the triggering dictionary [4, 9]. Hybrid methods combine statistical methods and pattern-based methods [12, 2].

Financial event subject extraction is a special case of event extraction. To the best of our knowledge, there is no method designed for this task in the literature. There are two relevant tasks, namely Named Entity Recognition (NER) [28] and Machine Reading Comprehension (MRC) [17]. NER is a classical NLP task which aims to recognize the entities like person, organization and location within a text. Typical methods of NER include stack-BiLSTMs [23], LM-LSTM-CRF [14] and GRN [3]. MRC is currently a hot research topic in NLP. It usually aims to answer questions from one or more relevant passages. Typical methods of MRC include Match-LSTM[25], BiDAF [20] and QANet [30]. Although MRC can be adapted to financial event subject extraction by treating event type as the question in MRC and event subject as the answer in MRC, no study on such an adaptation is reported in the literature. This paper proposes a novel adaptation of MRC to financial event subject extraction, which is proved to work well in our experiments.

3 Event-oriented Model with Focal Loss

The architecture of our proposed model is shown in Fig. 1. The input of our model is a pair composed of the given text and the given event type, which is then fed into a lexical encoder to generate the contextual representation of the given text. Afterwards, the trainable embeddings for event types are introduced to compute an attentive representation of the given text from the contextual representation. This attentive representation is then concatenated with the contextual representation to form an interactive representation of the given text. Finally, an answer pointer layer is employed to predict the start position and the end position of the event subject from the interactive representation.

3.1 Lexical Encoder

The input of the our proposed model is a pair (T, E), where $T = (w_1^T, ..., w_n^T)$ is a sequence of words representing the given text and $E = (w_1^E, ..., w_m^E)$ is a sequence of words representing the given event type. Following [7] we create a new sequence of tokens by first concatenating T and E and then adding a special token [CLS] in front of the first token w_1^T, a special token [SEP] between w_n^T and w_1^E, and another [SEP] behind the last token w_m^E. This new sequence of tokens

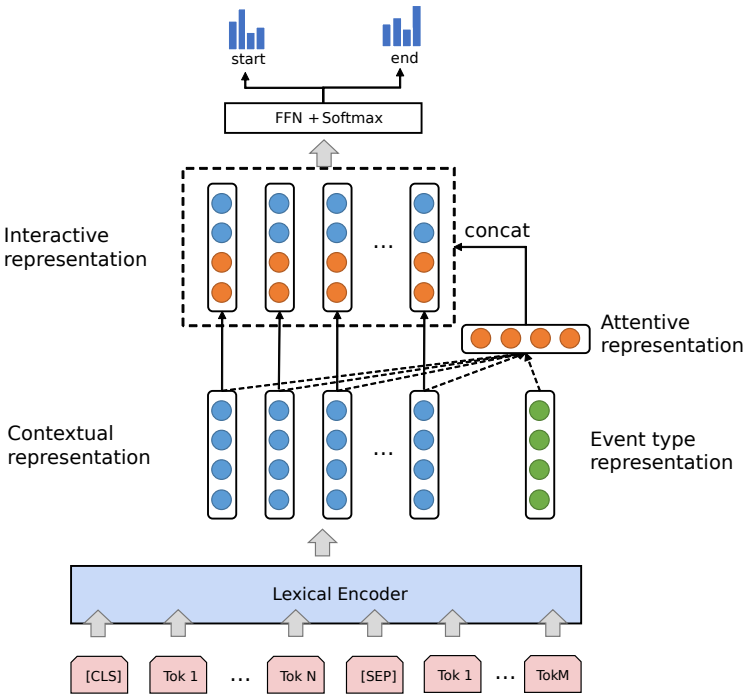

Fig. 1. The architecture of the proposed model.

is fed into a lexical encoder, which can be a pre-trained language model such as BERT [7], ERNIE [22] or BERT-wwm [6]. By $X \in \mathbb{R}^{t \times d}$ we denote the output of the lexical encoder, where t is the number of tokens in the new sequence and d is the dimension of a token embedding.

3.2 Interactive Composition Layer

To capture the interaction between the given text and the event type, we propose an interactive composition layer that introduces an attention mechanism over the contextual representation and the event type representation. Every event type is represented by an embedding that is a d-dimensional vector initialized randomly and tuned in the training course. The contextual representation and the event type representation are used to compute an attentive representation by a dot-product soft-attention mechanism. Formally, the attention value s_i for the i^{th} token in the input sequence of the lexical encoder is defined as

$$s_i = c_E^{\mathrm{T}} x_i \tag{1}$$

where c_E is the trainable d-dimensional embedding for E and x_i is the i^{th} vector in X the output matrix of the lexical encoder.

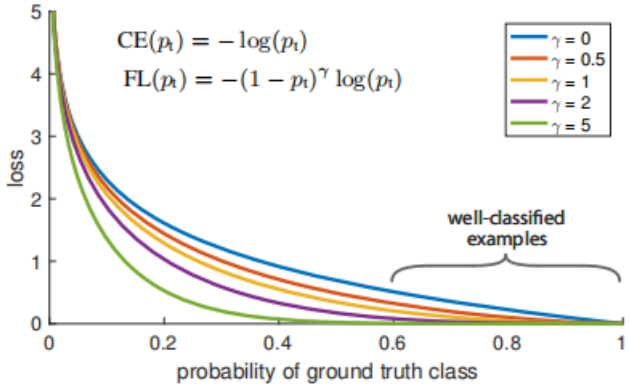

Fig. 2. The illustration of focal loss.

The attentive representation α for the input sequence of the lexical encoder is define as follows.

$$\alpha = \sum_{i=1}^{t} \frac{\exp(s_i)}{\sum_{j=1}^{t} \exp(s_j)} x_i \tag{2}$$

The interactive representation u_i for the i^{th} token in the input sequence of the lexical encoder is defined as

$$u_i = [x_i; \alpha] \tag{3}$$

where [;] denotes the concatenation function.

3.3 Answer Pointer Layer

The answer pointer layer is motivated by the pointer network [24] and is proposed in [25]. For extracting interactive features in a better way, a gate mechanism [21] with the Swish activation function [18] is employed to compress the concatenated vector u_i as follows.

$$g_i = W_1 u_i + b_1 \tag{4}$$

$$z_i = g_i \cdot \sigma(\beta g_i) \tag{5}$$

where $W_1 \in \mathbb{R}^{2d \times d}$ is a trainable matrix, b_1 is a trainable $2d$-dimensional vector, σ denotes the sigmoid activation function and β is a hyper-parameter.

Maxout networks [8] have been shown to be effective in accurately predicting the start position and the end position of an answer span [27]. Hence, in order to improve the prediction performance of the answer pointer layer, maxout networks are employed to compute a probability distribution on the start position and

a probability distribution on the end position for an answer span. Formally, the probability o_i^1 for ensuring the i^{th} token to be the start position and the probability o_i^2 for ensuring the i^{th} token to be the end position are defined by

$$h_i^1 = W_2 z_i + b_2 \qquad v_i^1 = \max_{j=1}^{l} h_{ij}^1 \tag{6}$$

$$h_i^2 = W_3 z_i + b_3 \qquad v_i^2 = \max_{j=1}^{l} h_{ij}^2 \tag{7}$$

$$o_i^1 = \frac{\exp(v_i^1)}{\sum_{j=1}^{t} \exp(v_j^1)} \qquad o_i^2 = \frac{\exp(v_i^2)}{\sum_{j=1}^{t} \exp(v_j^2)} \tag{8}$$

where l is a hyper-parameter in the maxout networks, $W_2 \in \mathbb{R}^{l \times 2d}$ and $W_3 \in \mathbb{R}^{l \times 2d}$ are trainable matrices, and b_2 and b_3 are trainable l-dimensional vectors.

3.4 Training with Three-stage Fine-tuning

To improve the prediction performance for minority event types, the above model is trained by using focal loss [13] in the minimization goal function and using a three-stage fine-tuning mechanism to dynamically set the values of the hyper-parameter γ of focal loss (see Fig. 2) in different epochs.

In the first stage, the hyper-parameter γ of focal loss is set as 0. This treatment amounts to fine-tuning the model by the traditional cross-entropy loss. Formally, the minimization goal function in this stage is defined as

$$\varepsilon_1(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log o_{y_i^1}^1 + \log o_{y_i^2}^2 \tag{9}$$

where θ denotes the set of all trainable parameters in the model, N is the number of training examples, y_i^1 and y_i^2 are the true start position and the true end position of the i^{th} example, respectively. Note that, when the i^{th} example has no event subject, both y_i^1 and y_i^2 are set as 1, corresponding to the [CLS] token.

In the second stage, the hyper-parameter γ of focal loss is increased to 1 to make the model focus more on training examples that are hard to be predicted correctly in the first stage. Formally, the minimization goal function in this stage is defined as follows.

$$\varepsilon_2(\theta) = -\frac{1}{N} \sum_{i=1}^{N} (1 - o_{y_i^1}^1) \log o_{y_i^1}^1 + (1 - o_{y_i^2}^2) \log o_{y_i^2}^2 \tag{10}$$

In the third stage, the hyper-parameter γ of focal loss is further increased to 2 to make the model focus on hard examples in a more forceful way. Formally, the minimization goal function in this stage is defined as follows.

$$\varepsilon_3(\theta) = -\frac{1}{N} \sum_{i=1}^{N} (1 - o_{y_i^1}^1)^2 \log o_{y_i^1}^1 + (1 - o_{y_i^2}^2)^2 \log o_{y_i^2}^2 \tag{11}$$

Since a pre-trained language model is usually fine-tuned in only a few epochs, we simply set every stage to consist in one epoch throughout our experiments.

Algorithm 1 The algorithm for answer refinement.

Require: The top-3 predicted answer spans $(s_i, e_i)_{1 \le i \le 3}$ with the highest probabilities $p_1 \ge p_2 \ge p_3$ for the given text $T = (w_1^T, \ldots, w_n^T)$ and the given event type E, and a probability margin δ, where p_i is the probability of (s_i, e_i) for $1 \le i \le 3$.

Initialize the resulting set of entities $\mathbb{D}$ as $\emptyset$
for each i from 1 to 3 such that $p_1 - p_i \le \delta$ **do**
  if the substring $(w_{s_i}^T, \ldots, w_{e_i}^T)$ contains a connection character "、" or "，" or "和" or "及" or "以及" **then**
    Set s_i' as the first position such that $s_i' < s_i$ and $w_{s_i'}^T$ is a connection character if there is a connection character before $w_{s_i}^T$, or as s_i otherwise
    Set e_i' as the last position such that $e_i' > e_i$ and $w_{e_i'}^T$ is a connection character if there is a connection character after $w_{e_i}^T$, or as e_i otherwise
    Split $(w_{s_i'}^T, \ldots, w_{e_i'}^T)$ by connection characters, yielding a set $\mathbb{U}$ of substrings
    Append $\mathbb{U}$ to $\mathbb{D}$
  else
    Add $(w_{s_i}^T, \ldots, w_{e_i}^T)$ to $\mathbb{D}$
  end if
end for
return $\mathbb{D}$

3.5 Answer Refinement

The aforementioned model can only predict answer spans together with their probabilities for extracting the event subject. But there can be multiple entities in an event subject for a given event type, thus we need to pick out the complete set of entities in the event subject from predicted answer spans. To this end we propose an algorithm for postprocessing the predicted answer spans, shown in Algorithm 1. In this algorithm, we only consider the top-3 predicted answer spans with the highest probabilities for the given text and the given event type, where the probability of an answer span (s, e) for s the start position and e the end position is defined as $(o_s^1 + o_e^2)/2$, and where o_s^1 and o_e^2 is defined by Equation (8). Among these top-3 predicted answer spans, we only handle the answer spans whose probabilities are not less than the probability of the top answer span by a user-specified margin δ. If the answer span being handled currently has a connection character "、", "，", "和", "及" or "以及", we widen this answer span to a new span from the first connection character before the span and the last connection character after the span, and then split the new span by connection characters. Every split substring is treated as an entity in the resulting event subject. Otherwise, we simply treat the answer span being handled currently as an entity in the resulting event subject.

4 Evaluation

In the CCKS 2019 challenge on financial event subject extraction, the organizers provided 17k labeled text and event type pairs for the training set, 3.5k unlabeled pairs for the validation set and 135k unlabeled pairs for the test set. All

Table 1. Performances of various methods on the validation set.

Model	F1-score(%)
BERT[7]	87.83
ERNIE[22]	88.26
BERT-wwm[6]	88.45
Our single BERT based model	88.93
Our single ERNIE based model	89.27
Our single BERT-wwm based model	89.46
Our ensemble model (3*ERNIE)	90.35
Our ensemble model (3*BERT-wwm)	90.72
Our ensemble model (all)	91.57
Our final model (with answer refinement)	92.58
Our final model on the test set	83.78

the evaluation results were carried out by an official evaluation system for this challenge[1], which outputs the classical F1-scores.

In our implementation of the proposed method, the lexical encoder was initialized by a pre-trained language model with 12 transformer layers which outputs 768-dimensional token embeddings, where all the transformer coders were built with 12 heads. We respectively tried three pre-trained language models BERT [7], ERNIE [22] and BERT-wwm [6]. The neural model shown in Fig. 1 was optimized by Adam with the warmup mechanism [7], where the initial learning rate was set as 5e-5, the warmup proportion as 10%, and the mini-batch size as 32. Besides training a single model, we also employed the ensemble strategy used in [26] to train three ensemble models, where the prediction probability vectors in these ensemble models was respectively averaged by three single models based on ERNIE (3*ERNIE), by three single models based on BERT-wwm (3*BERT-wwm), and by all these six single models (all). All the single models were trained with the same parameters. The hyper-parameter β used in Equation (5) was set as 1. The hyper-parameter l used in the maxout network namely Equations (6–7) was set as 50. The hyper-parameter δ used in Algorithm 1 was set as 0.45.

Table 1 reports the performances of various methods in terms of F1-score on the validation set as well as that of our final model (namely our ensemble models averaged by six single models and with answer refinement) on the test set. Our final model achieves the best performance among all the models. Specifically it achieves a 92.58% F1-score on the validation set and a 83.78% F1-score on the test set. We can also see that the pre-trained language model BERT-wwm outperforms both BERT and ERNIE. This superiority may be gained by the whole word masking mechanism [6] designed for Chinese. Our final model without answer refinement already improves the single model based on BERT-wwm by an absolute gain of 2.11% F1-score. Based upon this model, the answer refinement postprocessing step contributes to a further absolute gain of 1.01% F1-score.

[1] https://www.biendata.com/competition/ccks_2019_4/

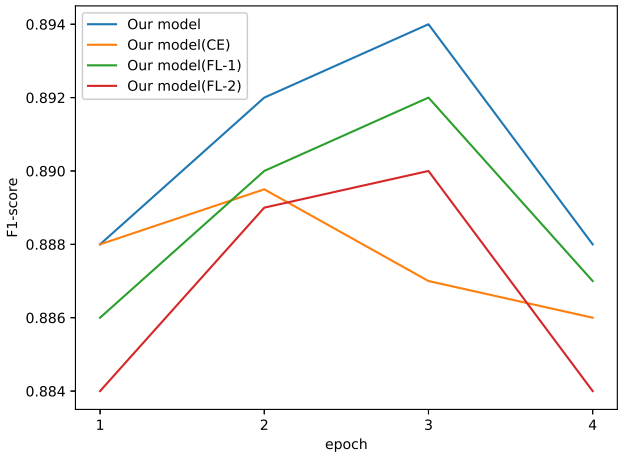

Fig. 3. The F1-score on the validation set of the proposed model with different losses.

Fig. 3 shows the F1-scores of our single BERT-wwm based model with different hyper-parameters γ in focal loss. The blue line corresponds to our model with the proposed three-stage fine-tuning mechanism, namely with increasing values of γ in different epochs, where in the additional (i.e. the 4th) epoch γ is set as 2. The orange line corresponds to our model (CE) with γ fixed to 0, namely using the cross-entropy loss, in all epochs. The green line corresponds to our model (FL-1) with γ fixed to 1 in all epochs. The red line corresponds to our model (FL-2) with γ fixed to 2 in all epochs. It can be seen that the three-stage fine-tuning mechanism outperforms other variants in terms of F1-score.

5 Conclusions

In this paper we have proposed a neural model based method for financial event subject extraction. The neural model builds upon a pre-trained language model such as BERT, by adding an interactive composition layer that exploits an attention mechanism to capture the dependence between event subjects and event types, and by further adding an answer pointer layer to predict answer spans. The training course is also enhanced by a three-stage fine-tuning mechanism with focal loss. In addition, an answer refinement algorithm is proposed to extract event subject entities from predicted answer spans. Experimental results on the CCKS 2019 challenge show that the proposed method achieves a 92.58% F1-score on the validation set and a 83.78% F1-score on the test set.

Acknowledgements

This work is partly supported by National Natural Science Foundation of China (61876204) and Science and Technology Program of Guangzhou (201804010496).

References

1. Ahn, D.: The stages of event extraction. In: The Workshop on Annotating and Reasoning about Time and Events. pp. 1–8 (2006)
2. Björne, J., Ginter, F., Pyysalo, S., Tsujii, J., Salakoski, T.: Complex event extraction at pubmed scale. Bioinformatics [ISMB] **26**(12), 382–390 (2010)
3. Chen, H., Lin, Z., Ding, G., Lou, J., Zhang, Y., Karlsson, B.: GRN: gated relation network to enhance convolutional neural network for named entity recognition. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 6236–6243 (2019)
4. Chen, Y., Liu, S., Zhang, X., Liu, K., Zhao, J.: Automatically labeled data generation for large scale event extraction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 409–419 (2017)
5. Chen, Y., Xu, L., Liu, K., Zeng, D., Zhao, J.: Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of the 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, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. pp. 167–176 (2015)
6. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-training with whole word masking for chinese BERT. CoRR **abs/1906.08101** (2019)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
8. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A.C., Bengio, Y.: Maxout networks. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013. pp. 1319–1327 (2013)
9. Gu, J., Lu, Z., Li, H., Li, V.O.K.: Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers (2016)
10. Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F., Caron, E.: A survey of event extraction methods from text for decision support systems. Decision Support Systems **85**, 12–22 (2016)
11. Ji, H., Grishman, R.: Refining event extraction through cross-document inference. In: ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA. pp. 254–262 (2008)
12. Jungermann, F., Morik, K.: Enhanced services for targeted information retrieval by event extraction and data mining. In: LWA 2008 - Workshop-Woche: Lernen, Wissen & Adaptivität, Würzburg, Deutschland, 6.-8. Oktober 2008, Proceedings. pp. 50–56 (2008)
13. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2999–3007 (2017)

14. Liu, L., Shang, J., Ren, X., Xu, F.F., Gui, H., Peng, J., Han, J.: Empower sequence labeling with task-aware neural language model. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 5253–5260 (2018)
15. Liu, S., Chen, Y., Liu, K., Zhao, J.: Exploiting argument information to improve event detection via supervised attention mechanisms. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 1789–1798 (2017)
16. Nguyen, T.H., Cho, K., Grishman, R.: Joint event extraction via recurrent neural networks. In: NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. pp. 300–309 (2016)
17. Qiu, B., Chen, X., Xu, J., Sun, Y.: A survey on neural machine reading comprehension. CoRR **abs/1906.03824** (2019), http://arxiv.org/abs/1906.03824
18. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings (2018)
19. Reichart, R., Barzilay, R.: Multi-event extraction guided by global constraints. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada. pp. 70–79 (2012)
20. Seo, M.J., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)
21. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. CoRR **abs/1505.00387** (2015)
22. Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., Wu, H.: ERNIE: enhanced representation through knowledge integration. CoRR **abs/1904.09223** (2019)
23. Tran, Q., MacKinlay, A., Jimeno-Yepes, A.: Named entity recognition with stack residual LSTM and trainable bias decoding. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers. pp. 566–575 (2017)
24. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 2692–2700 (2015)
25. Wang, S., Jiang, J.: Machine comprehension using match-lstm and answer pointer. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)
26. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 4144–4150 (2017)
27. Xiong, C., Zhong, V., Socher, R.: Dynamic coattention networks for question answering. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)

28. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. pp. 2145–2158 (2018)
29. Yang, H., Chen, Y., Liu, K., Xiao, Y., Zhao, J.: DCFEE: A document-level chinese financial event extraction system based on automatically labeled training data. In: Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations. pp. 50–55 (2018)
30. Yu, A.W., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M., Le, Q.V.: Qanet: Combining local convolution with global self-attention for reading comprehension. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (2018)