Event Causality Identification by Modeling Events and Relation Embedding

Zhenyu Yang, Wei Liu(✉), and Zongtian Liu

School of Computer Engineering and Science,
Shanghai University, Shanghai 200444, China
18201789699@163.com, {liuw,ztliu}@shu.edu.cn

Abstract. Events and event relations contain high-level semantic information behind texts. In this paper, we mainly discuss event causality relation identification. Traditional approaches of causality relation identification rely on the recognition of casual relationship connectives or manual features of causality relationships, and these methods have disadvantage of low recognition coverage and being lack of adaptive. To solve this problem, we propose a novel model based on modeling event and event relation. We use word sequence around event trigger as input data and use event based Siamese Bi-LSTM network to model events by encoding the event representations into a fixed size vectors, and then these events representations are applied in relation embedding training and prediction. Experimental results show that the proposed method can achieve better effect on CEC 2.0 corpus.

Keywords: Siamese network · Event relation · LSTM · CEC

1 Introduction

Natural language organized texts express higher-level semantic information through events. Recognizing these events and the relationships between these events can help computers easily understand the precise meaning of texts and lay a solid foundation for the reasoning and modeling of event ontology.

We define an event as a thing happens in a certain period of time and place, in which some actors participate and show some features of action, also accompany with the changing of internal status [1]. An event trigger is the word that most exactly expresses the occurrence of an event. For example: in the sentence "the earthquake happened yesterday caused 21 wounded". "wounded" is a trigger of event. Event trigger is the most significant signal of event in texts.

Event can be formalized as a 6-tuple $e = (A, O, T, P, S, L)$. We call elements in 6-tuple event elements, and represent action, object, time, place, status, language expression respectively. In natural language processing, we mainly focus on participants, objects, time, and location of an event. These elements present as word in natural language and contains important information of events.

Causality relation is a kind of common and important relation between events. If an event e_1 happened, the another event e_2 happens with the probability above the threshold of causality, there is a causality relation between e_1 and e_2. Causality relation

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11303, pp. 59–68, 2018.
https://doi.org/10.1007/978-3-030-04182-3_6

can be divided into explicit causality and implicit causality. Explicit causality denotes those relations exist connectives exactly express the relation between events. Implicit causality denotes those relations lack exact connectives and need to be speculated by the contexts. In addition, there're three relations between events beside causality relation, which include composition relation, follow relation and concurrency relation. If an event *e* can be decomposed to several sub-events e_i with smaller granularity, there exists composition relation between *e* and e_i. If in a certain length of time, the occurrence of event e_1 follows the occurrence of the event e_2 at above specified threshold, there exists a follow relation between e_1 and e_2. If there are event e_1 and event e_2 occur simultaneously in a certain period of time, there is a concurrency relation between e_1 and e_2.

Current researches on causality relation identification are mostly based on the feature selection, pattern matching and rule reasoning. These approaches of causality relation identification can't realize the context and identify the implicit causality relation in texts.

In recent years, deep learning (DL) within the machine learning field has shown that it can be successfully applied to reduce the data dimension by extracting deep features of data and use those features to present better results than traditional machine learning methods. Although there are preliminary applications of DL in many natural language processing (NLP) tasks. There are few researches on causality relation identification based on DL. Therefore, we propose a new method based on Siamese network. Firstly we use Bi-LSTM network to capture the semantic information in events and generate event representations which cover event elements and event triggers. Then we use the element-wise difference between events to predict the causality relation. The experimental results show that our proposed model has achieved better performance in causality relation identification. In addition, event representations generated by our proposed model also achieve satisfactory results in the task of event classification.

The remained of this paper is organized as follows: we describe the related works in Sect. 2. Our proposed model is described in Sect. 3. Section 4 presents our experimental results. Finally, we conclude in Sect. 5.

2 Related Work

2.1 Siamese Network

Siamese network is a special type of neural network architecture which is widely applied in calculating the similarity of pair of inputs like texts or pictures [2–4]. Siamese network proposed by Chopra consists of two identical neural networks with shared parameters and the last layers of two networks are then fed to a contrastive loss function which calculates the similarity between two inputs. Chopra's work illustrates the method for learning complex similarity metrics with a face verification application. Recently, Siamese Network is also applied in NLP. Kenter [5] presented the Siamese CBOW model based on Siamese Network. Siamese CBOW handles the task of sentence representation by training word embedding directly, and then trains a sentence embedding by predicting from its surrounding sentence representations. Muller [6] proposed their Manhattan LSTM (MaLSTM) for assessing the semantic similarity metric between sentences. The work demonstrates that a simple LSTM is capable of modeling complex semantics if the representations are explicitly guided.

2.2 Causality Relation Identification

Broadly speaking, causality relation identification refers to the method of knowing whether an event causes another. By analyzing the verbs that express causality relation in French, Garcia [7] proposed a COATIS system to extract the explicit causality relation in French. Khoo [8] proposed an automatic method for identifying causality relation in Wall Street Journal text using linguistic clues and pattern-matching. Girju [9] searched for causal verbs through the Internet and WordNet to establish the Lexico-syntactic model, which enables automatic recognition of causality relations for specific events.

However these methods based on pattern-matching are domain-specific and require a lot of artificial markings. Therefore, recent studies have used methods based on machine learning and statistical probabilities to identify causality relation.

For example, Marco [10] adopted the Naive Bayesian to identify explicit causality relation by analyzing the probabilities of words between adjacent sentences. Inui [11] used support vector machine (SVM) to identify explicit causality relation in corpus by using the specific language components between the indicator and the sentence. Zhong [12] proposed a method based on cascaded model to identify explicit causality relation.

Although methods above work well, they are limited to the identification of explicit causality relation. In fact, there're a lot of implicit causality relations in texts. Therefore, there are also researchers who have studied the identification of implicit causality relation.

Fu [13] casted the causality relation identification as event sequence labeling and proposed dual-layers CRFs model to label the causal relation of event sequence. Yang [14] proposed correlation degree RCE to describe the probabilities between events and set threshold as a binary prediction to predict an event pair as causality or not.

The researches of causality relation identification above are mostly based on the feature selection, pattern matching and rule reasoning. Some scholars pay attention to the causality connectives rather than the relation between semantic information of events. In this paper, we propose a method to generate event representations based on event trigger and event elements. Event representations are used to predict the causality relation between events.

3 Proposed Model

3.1 Structure of Proposed Model

Researchers in the field of Knowledge Graph (KG) embed knowledge graph components (entities and relations) in continuous vector space while preserving properties of the original data, such as TransE [15], TransH [16] and TransD [17]. In TransE, relations are represented as translation embedding in vector space, if a triplet (subject, relation, object) exists in KG, we want that object should be close to subject + relation, while subject + relation should be far away from object if the triplet doesn't exist. Once the model has learned an embedding vector for each entity and relation, predictions will be performed by using the same translation approach in embedding space. For example, the prediction for a given subject-relation is generated by searching for the nearest neighbor entity of subject + relation in vector space.

In the field of event-oriented knowledge representation, events and event relations can be considered as special entities and relations. If we use certain methods to represent events and event relations in continuous vector space, we can also predict the relation type between events.

Based on the ideas above, this paper proposes our proposed model based on Siamese Architecture shown in Fig. 1. There are two networks Bi-LSTM$_a$ and Bi-LSTM$_b$ which each processes one of the events in a given pair and they share parameters. We use Bi-directional long short time memory (Bi-LSTM) networks to obtain event representations. Then event representations generated by Siamese LSTM Network are used to train relation embedding.

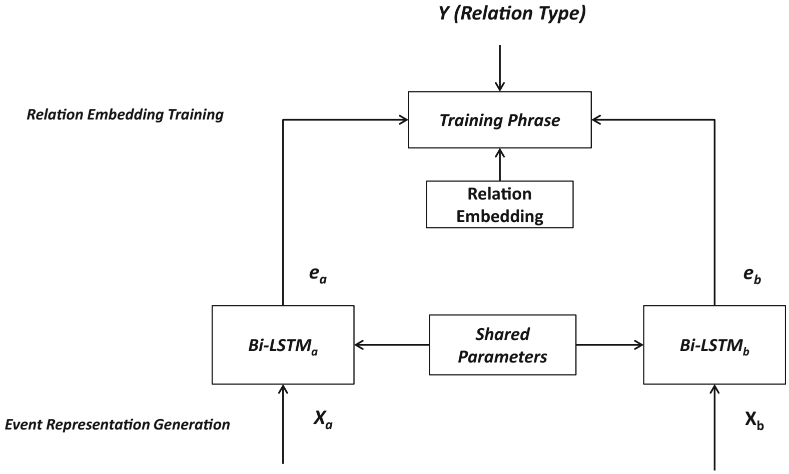

Fig. 1. The training process of proposed model

3.2 Event Representation Generation

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing where words or phrases from the vocabulary are mapped to vectors of real numbers. Word embedding proposed by Mikolov [18, 19] can be trained to capture semantic and syntactic relationships between words, by mapping related words to vectors that lie close in the embedding vector space. In summary, word embedding provides us an efficient method to represent word in vector space. In this paper, pre-trained word embedding is used to convert words into dense vectors.

In order to represent event, we introduce a sequence model Recurrent Neural Network (RNN). RNN is a powerful model for learning features from sequential data. RNN model is suitable for our inputs which are sequences of words, and since neural networks receive fixed size vectors or matrixes as input, words are converted into word embedding before used as inputs. Bi-directional RNN (Bi-RNN) uses a finite sequence to model sequence based on past and future contexts. This is done by concatenating the hidden states of two RNN, one processing the sequence from left to right, the other one from right to left. We can update the hidden state of each timestamp t as following:

$$h_{ft} = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{1}$$

$$h_{bt} = \sigma(W_b h_{t+1} + U_b x_t + b_b) \tag{2}$$

$$h_t = h_{ft} \oplus h_{bt} \tag{3}$$

In formulas above, h_{ft} is the hidden state of timestamp t along the forward direction (from left to right), h_{bt} is the hidden state of timestamp t along the backward direction (from right to left), h_{t} is the hidden state at timestamp t and $\oplus$ denotes the concatenating operation between two vectors.

Although RNNs present acceptable performance in sequences processing, the optimization of the weight matrixes is difficult because its backpropagated gradients vanish over long sequences. LSTM networks are introduced to avoid the long-term dependency problem. Like RNNs, LSTM sequentially updates a hidden-state representation.

In this paper, we use Bi-RNNs with LSTM cell which is called Bi-LSTM and introduced above to learn event representation. The learning process is shown in Fig. 2.

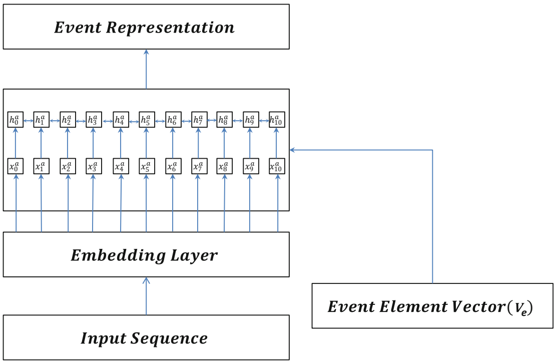
Event Representation
h_0^a h_1^a h_2^a h_3^a h_4^a h_5^a h_6^a h_7^a h_8^a h_9^a h_{10}^a
x_0^a x_1^a x_2^a x_3^a x_4^a x_5^a x_6^a x_7^a x_8^a x_9^a x_{10}^a
Embedding Layer
Input Sequence
Event Element Vector(V_e)

Fig. 2. The training process of proposed model

In this paper, word sequences with fixed length are used as input to represent events. Word sequences contain five words behind event triggers, event trigger word and five words after event triggers in texts. We use "<pad>" to represent paddings in word sequences which make length of input sequence equal. In CEC, we find that the average distance between event triggers and event element(such as time, place and object) are 3.4 and 96% of event elements can be covered when the length of word sequence is eleven. So we set the length of word sequence as eleven. Firstly, word sequence is converted to dense word embedding by embedding layer, and then input into Bi-LSTM model. After the processing of Bi-LSTM model, we finally get hidden set H = $\{h_0, h_{1,}.. h_{10}\}$. The event representation e can be obtained by following formula. Where f_t represents the feature of event trigger and f_e represents the feature of event elements. We use hidden state h_5 which is the hidden state of event trigger to represent feature of event trigger f_t. When $\alpha = 0$ event representation e excludes the feature of event elements.

$$e = (1 - \alpha) * f_t + \alpha * f_e \tag{4}$$

In addition to the feature of event trigger, our event representation also focuses on feature of event elements. One-hot vector v_e is used to denote whether the word in timestamp i of input sequence is event elements. Feature of event elements can be obtained as following:

$$f_e = \frac{1}{\sum_{i=0}^{L} v_{e_i}} \sum\nolimits_{i=0}^{L} v_{e_i} * h_i \tag{5}$$

Where $v_{ei} \in \{0,1\}$ is the value in i-dimension of v_e, h_i is the hidden state in timestamp i generated by Bi-RNN model which is discussed in the above, L is the length of input sequence.

3.3 Training Relation Embedding

Given a training set S of triplets (e1, e2, r) composed of two events e1, e2 and a relation $r \in R$, our model learns the representations of events and relations. The basic idea in this step is minimize the Dist(e1, e2, r) for each training example. Dist(e1, e2, r) is calculated as following:

$$Dist(e1, e2, r) = ||e1 + r - e2|| \tag{6}$$

In the task of relation identification, we introduce the loss function as following, where c is a constant, r_{pos} is the relation between e_1 and e_2 r_{neg} represent the negative relation between e_1 and e_2. The second item on the right of the equation is the training example, while the third item is the corrupted example we generated in order to make $e_1 - e_2$ be away from the corrupted event relation.

$$loss = c - Dist(e1, e2, r_{pos}) + Dist(e1, e2, r_{neg}) \tag{7}$$

r_{neg} is calculated as following:

$$r_{neg} = \frac{1}{N-1} \sum\nolimits_{r \in R - \{r_{pos}\}} r \quad (8)$$

4 Experimental Result

4.1 Experiment Dataset

Our experimental dataset is CEC 2.0. CEC 2.0 is an event-based Chinese natural language corpus developed by the Semantic Intelligence Laboratory of Shanghai University. It has collected 333 newspaper reports about earthquakes, fires, traffic accidents, terrorist attacks and food poisoning. We labeled event triggers, participants, objects, times, places and relationships between events by using a semi-automatic method. Statistics of events and relationships labeled exactly is shown in Table 1.

Table 1. Statistics of event types and event relation

Event type	Amount	Event relation type	Amount
Perception	264	Follow relation	702
StateChange	996	Causality relation	806
Emergency	667	Concurrency relation	504
Statement	859	Overall	2008
Action	1121		
Operation	1245		
Movement	469		
Overall	5621		

4.2 Event Causality Identification

We compare our proposed model's results with other models shown in Table 2. Yang [14] defined causal correction degree (RCE) to predict whether causality exists between events. Zhong [12] proposed a cascaded model based on the bootstrapping algorithm to identify causality relation. Girju's method [9] is based on pattern-matching. From the results, we find absolute increment when α increases, and the highest F-Measure is 83.82%. At the same time, we also notice that the performance decline when $\alpha > 0.2$ and proposed model ($\alpha = 0.5$) even achieves worse result than proposed model ($\alpha = 0$). The result demonstrates that the feature of event elements really work in the event representations and enrich the semantic information of the event. However, if the model focuses on event elements excessively, important information will be ignored. Compared with other models, Proposed model ($\alpha = 0.2$) has shown slight improvement in F-Measure. The proposed model's ability to capture the semantic information of the event is likely to be one of the reasons of improvement in performance.

Table 2. Performance Comparison of all models in causality relation identification

Method	Precision (%)	Recall (%)	F-Measure (%)
Yang's method [14]	62.20	58.00	59.90
Zhong's method [12]	**85.39**	77.53	81.27
Girju's method [9]	73.91	**88.69**	80.63
Proposed model ($\alpha = 0$)	79.01	80.34	79.67
Proposed model ($\alpha = 0.1$)	82.07	81.16	81.61
Proposed model ($\alpha = 0.2$)	83.01	84.65	**83.82**
Proposed model ($\alpha = 0.3$)	82.51	81.62	82.07
Proposed model ($\alpha = 0.4$)	82.63	79.29	80.93
Proposed model ($\alpha = 0.5$)	77.04	79.89	78.44

4.3 Event Recognition

In this paper, we apply Bi-LSTM network in proposed model to learn event representation which can represent the content of events. To evaluate the practicality of our representations of events generated in our proposed model, we applied the Bi-LSTM network trained for the task of event relation identification into the task of event classification. We use SVM classifier to classify the events in CEC.

We also compare our proposed model's results with other models proposed for the task of event classification shown in Table 3. Fu et al. [20] proposed classifier based on SVM and dependency parsing. Zhao et al. [21] proposed a classifier based on maxium entropy with defined features. Our proposed model exactly capture context information of events, and the event embeddings perform well in the task of event classification.

Table 3. Performance Comparison with related works in event classification

Method	Precision (%)	Recall (%)	F-Measure (%)
Event representation generated by proposed model + SVM classifier	**81.10**	**81.16**	**81.01**
Fu's method [20]	71.60	67.20	69.30
Zhao's method [21]	57.14	64.22	60.48

5 Discussion and Conclusion

This paper presented a novel method for event causality relation identification based on modeling events and relations on dense vector space. We use word sequence around event triggers as input and learn event embedding by Siamese Bi-LSTM network in relation identification task. The Bi-LSTM learns the features of event trigger and event elements. Experimental results show that our method achieves good performance and the best F-Measure of the causality relation arrives at 83.82%. Furthermore, we applied Bi-LSTM network trained in relation identification to generate event representations and use them in event classification task. The results show that event representations perform very well and our proposed model really capture important context information of events.

In future work, we will improve the performance and scalability of proposed model, meanwhile we will try to apply the approach in proposed model in event reasoning and find out more semantic information behind events and relations and dig out more event knowledge for event-based natural language processing.

References

1. Liu, Z.T., et al.: Research on event-oriented ontology model. Comput. Sci. **36**(11), 189–192 (2009)
2. Chopra, S., Hadsell, R., Lecun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), pp. 539–546. IEEE, Piscataway (2005)
3. Norouzi, M., Fleet, D.J., Salakhutdinov, R.R.: Hamming distance metric learning. In: Advances in Neural Information Processing Systems, pp. 1061–1069. Curran Associates, New York (2012)
4. Baraldi L., Grana C., Cucchiara R.: A deep siamese network for scene detection in broadcast videos. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1199–1202. ACM, New York (2015)
5. Kenter, T., Borisov, A., de Rijke, M.: Siamese CBOW: optimizing word embeddings for sentence representations (2016). arXiv preprint: arXiv:1606.04640
6. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: Thirtieth AAAI Conference on Artificial Intelligence, pp. 2786–2792. AAAI, Menlo Park (2016)
7. Garcia, D.: COATIS, an NLP system to locate expressions of actions connected by causality links. In: Plaza, E., Benjamins, R. (eds.) EKAW 1997. LNCS, vol. 1319, pp. 347–352. Springer, Heidelberg (1997). https://doi.org/10.1007/BFb0026799
8. Khoo, C.S.G., Kornfilt, J., Oddy, R.N., Myaeng, S.H.: Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary Linguist. Comput. **13**(4), 177–186 (1998)
9. Girju, R.: Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, vol. 12, pp. 76–83. ACL, Stroudsburg (2003)
10. Marcu, D., Echihabi, A.: An unsupervised approach to recognizing discourse relations. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 368–375. ACL, Stroudsburg (2002)
11. Inui, T., Inui, K., Matsumoto, Y.: What kinds and amounts of causal knowledge can be acquired from text by using connective markers as clues? In: Grieser, G., Tanaka, Y., Yamamoto, A. (eds.) DS 2003. LNCS (LNAI), vol. 2843, pp. 180–193. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39644-4_16
12. Zhong, J., Long, Y., Tian, S.: Causal relation extraction of uyghur emergency events based on cascaded model. Zidonghua Xuebao/Acta Autom. Sin. **40**(4), 771–779 (2014)
13. Fu, J., Liu, Z., Liu, W.: Using dual-layer CRFs for event causal relation extraction. IEICE Electron. Express **8**(5), 306–310 (2011)
14. Yang, J., Liu, Z., Liu, W.: Identify causality relationships based on semantic event. J. Chin. Comput. Syst. **36**(3), 433–437 (2016)

15. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: International Conference on Neural Information Processing Systems, pp. 2787–2795. Curran Associates, New York (2013)
16. Wang, Z., Zhang, J., Feng J.: Knowledge graph embedding by translating on hyperplanes. In: AAAI - Association for the Advancement of Artificial Intelligence, pp. 1112–1119. AAAI, Menlo Park (2014)
17. Ji, G., He, S., Xu, L.: Knowledge graph embedding via dynamic mapping matrix. In: Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, pp. 687–696. ACL, Stroudsburg (2015)
18. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196. JMLR (2014)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). arXiv preprint: arXiv:1301.3781
20. Fu, J., Liu, Z., Zhong, Z.: Chinese event extraction based on feature weighting. Inf. Technol. J. **9**(1), 184–187 (2010)
21. Zhao, Y., Qin, B., Che, W., Liu, T.: Research on Chinese event extraction. J. Chin. Inf. Process. **22**(1), 3–8 (2008)