A KNOWLEDGE/DATA ENHANCED METHOD FOR JOINT EVENT AND TEMPORAL RELATION EXTRACTION

Xiaobin Zhang[1,2]*,Liangjun Zang*[1✉] *,Peng Cheng*[3]*,Yuqi Wang*[3]*,Songlin Hu*[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]State Key Laboratory of Media Convergence Production Technology and Systems, Beijing, China

ABSTRACT

Understanding temporal relations (TempRels) between events is an important task that could benefit many downstream NLP applications. This task inevitably faces the challenges of both a limited amount of high-quality training data and a very biased distribution of TempRels. These problems will substantially hurt the performance of extraction systems because they are inclined to predict dominant TempRels when training with a limited amount of data. To alleviate those issues, we propose a Knowledge/Data Enhanced method for Event and TempRel Extraction, which integrates the temporal commonsense knowledge, data augmentation and Focal Loss function into one single extraction system. Altogether, these components improve the performance of the system on two public benchmark datasets TB-Dense and MATRES[1].

Index Terms— Event Extraction, Temporal Relation, Commonsense Knowledge, Data Augmentation

1. INTRODUCTION

Understanding temporal relation between events is an important task that benefits many downstream NLP applications such as story generation, summarization, timeline construction[1, 2]. Traditional statistical methods use manually designed features such as temporal connectives or modal verbs to recognize temporal relations between events[3]. Recently, deep learning model have been widely used to learn representations of events and temporal relations and achieve moderate improvements[4, 5],While these models have made remarkable progress, their performance were limited by the corpus they used. The first problem is the lack of large-scale training data, since the annotation process of events and temporal relations is very labor-intensive and has low inter-annotator agreements(IAA)[1, 6]as pointed in TimeBank [7] and TempEval3[1]. The second one is the label imbalance problem, i.e., the numbers of samples of different relations vary greatly in the training set, which makes the performance on few-shot relations much worse than the others.

To eliminate the limitations of small-scale imbalanced training data, researchers proposed new methods for improving temporal relation extraction with external knowledge resources. Intuitively, if the context between two event mention words has neither any temporal connective (e.g., before, after, and since) nor any modal verb (e.g., will, would), people can still recognize their temporal relations using common sense knowledge. Therefore, Ning et al.[2] integrate a LSTM network with a Common Sense Encoder (CSE), where the CSE is a Siamese network trained on a commonsense knowledge base called TEMPROB[8]. Han et al.[5]formulate the incorporation of probabilistic knowledge TEMPROB as a constrained inference problem, and use it to improve the outcomes from neutral models. However, external knowledge may introduce new noises to the system and it is unclear under what conditions to use this knowledge. In addition, the knowledge does not cover 'include' or 'simultaneous' relations since such knowledge is not easy to obtain.

In this paper we propose a new Knowledge/Data Enhanced framework of Extracting Events and Temporal Relations. Firstly, we introduce the prior distribution knowledge TEMPPROB that events often follow, and decide whether to use this knowledge with selective attention. Secondly, we generate new samples for temporal relations by exchanging event pairs of the same relation between samples in the training data. Finally, we use Focal Loss as training objective, which makes our model focus on the few-shot relations that is hard to train. Altogether, these components improve the system performance, and they could be readily applied to any system of extracting events and TempRels. Experimental results on two public benchmark datasets show that our system improves over BiLSTM+MAP [4] by 1.09% in F1 on TB-Dense dataset and by 1.12% in F1 on MATRES dataset. The proposed system is public and can serve as a strong baseline for future research.

Our contribution is summarized as follows: Firstly, we propose a knowledge/data enhanced method for extracting events and temporal relations, which integrates external knowledge, data augment and focal loss objectives together. Secondly, these components altogether improve the TempRel

[1]We release our code and dataset at https://github.com/buaadk/KJETE/

extraction system and achieve better performances on two public benchmark datasets.

2. METHODOLOGY

In this section we first provide an overview of our neural model, and then describe each component in our framework in detail (i.e., external knowledge fusion, data augment and FocalLoss).

2.1. Overall Architecture

Given a raw text as input, our goal is to identifies all events and classifies temporal relations for all predicted event pairs. We denotes the set of all possible events as $\mathcal{E}$, the set of all possible event pairs as $\mathcal{EE}$, and the set of all possible temporal relations as $\mathcal{R}$.

The architecture of our model is shown in Fig.1. We employ the pre-trained model RoBERTa [9][2] to produce the contextualized embeddings and use BiLSTM network for event and TempRel prediction[3]. We integrate prior knowledge into the contextual representation of each event pair candidates, and then use a feed-forward neural network(FFNN) to generate confidence scores for events and TempRels. In the prediction layer, we use softmax function to output the probability distribution of event and TempRels.

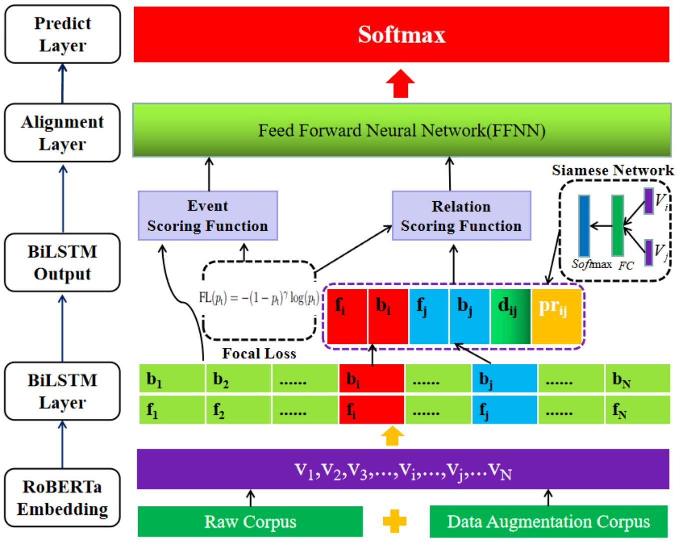

Predict Layer
Alignment Layer
BiLSTM Output
BiLSTM Layer
RoBERTa Embedding
Softmax
Feed Forward Neural Network(FFNN)
Siamese Network
Event Scoring Function
Relation Scoring Function
V_i
V_j
Softmax
FC
$FL(p_t) = -(1-p_t)^\gamma \log(p_t)$
Focal Loss
f_i
b_i
f_j
b_j
d_{ij}
pr_{ij}
b_1 b_2 ...... b_i ...... b_j ...... b_N
f_1 f_2 ...... f_i ...... f_j ...... f_N
$V_1,V_2,V_3,\ldots,V_i,\ldots,V_j,\ldots V_N$
Raw Corpus
Data Augmentation Corpus

Fig. 1. The overview of model structure.

2.2. External Knowledge Fusion

We follow the method [2] to produce the prior probability of temporal relation of each event pair. Specifically, we introduce the TEMPROB resource[8], which contains observed frequencies of event pairs following specific temporal order from a large corpus, and use it to train a Siamese network as the common sense encoder (CSE). The CSE provides the prior probability p_r of a given temporal relation r that usually follows by two events (i, j).

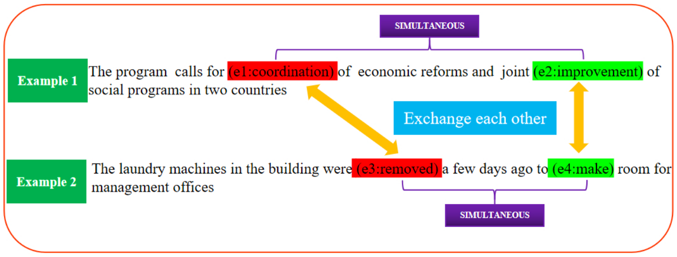
Intuitively, we usually learn linguistic and semantic information from the context around event expressions in order to decide whether we use our commonsense knowledge. For example, if the trigger words of two events are far apart in text, it is more likely to use commonsense knowledge to infer their temporal relation. Hence, we define the probability of a given temporal relation r between two events e_i and e_j given its prior knowledge p_r as follows:

$$p(r|e_i, e_j, p_r) = \frac{exp(p_r \cdot A \cdot (f_i \oplus b_i \oplus f_j \oplus b_j \oplus d_{i,j}))}{\sum_{k=1}^{|\mathcal{R}|} exp(p_r \cdot A \cdot (f_i \oplus b_i \oplus f_j \oplus b_j \oplus d_{i,j}))} \tag{1}$$

where the A is a diagonal matrix, f_i and f_j are the BiLSTM layer forward outputs, b_i and b_j represent the BiLSTM layer backward output, and $d_{i,j}$ is the manually designed linguistic feature vector.

2.3. Data Augmentation

To generate more samples for the few-shot temporal relations such as INCLUDES, ISINCLUDED and SIMULTANEOUS, we introduce data augmentation method to our system. An example of generating new instance of SIMULTANEOUS relation is shown in the Fig.2



The program calls for (e1:coordination) of economic reforms and joint (e2:improvement) of social programs in two countries

The laundry machines in the building were (e3:removed) a few days ago to (e4:make) room for management offices

Fig. 2. A example of data augmentation.

1. We select the sentences that contain any few-shot TempRel label from the corpus.

2. If two sentences have the same label, we first swap their event trigger words, and then randomly mask and fill some words in contexts with the pre-trained language model RoBERTa. Thus, we get two new sentences that contains the TempRel.

[2] We employed the officially published standard RoBERTa-Large model with 1024 hidden sizes,24 layers,16 heads implemented by https://github.com/pytorch/fairseq/tree/main/examples/roberta/

[3] Following the previous work of event prediction [3, 10, 4, 5],we extract events from the current sentence and the neighboring sentences

3. From all generated sentences, we select some high-quality sentences according to their confidence scores of the model BiLSTM+MAP[4], and add them to the training data.

2.4. Focal Loss

To address the label imbalance issue, we introduced the Focal Loss[11] to our loss function. According to the author, if the number of easily classified samples is too large, they will be the majority of the total loss. Thus, it is necessary to integrate the Focal Loss function into our model. The definition of focal loss is as follows:

$$FL(p_t, \theta) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \tag{2}$$

where α_t is used to balance the weight of each label, p_t denote the probability of predicted label, and $(1 - p_t)^{\gamma}$ is the modulating factor. Here we set the tunable focusing parameterγ to 1.

2.5. Training Objectives

Our objective function is defined as follows:

$$\mathcal{J}(\theta_1, \theta_2) = \alpha \sum_{i=1}^{|\mathcal{E}|} FL(e_i, \theta_1) + \beta \sum_{j=1}^{|\mathcal{EE}|} FL(r_j, \theta_2) + \gamma ||\Phi||^2 \tag{3}$$

where $FL(y_i, \theta_1)$ is the event predict loss, $FL(r_j, \theta_2)$ is the temporal predict loss, $||\Phi||^2$ is the L2 norm regularization term. Here $|\mathcal{E}|$ is the number of candidate events, $|\mathcal{EE}|$ is the number of candidate event pairs, e_i denotes the true label of event trigger, r_j denotes the true label of true TempRel. In addition, α, β and γ are the hyper-parameters to balance the losses between event, relation and the regularizer, where we set $\alpha = 10$, $\beta = 1$ and $\gamma = 1$ and discuss them in detail in a later section.

2.6. Implementation Details

We introduce hyper-parameters in this subsection. We initialize the embedding size and hidden size with 100 dimensions. Our model is trained using Adam optimizer, where the learning rate is set to 1e-3, the dropout rate is set to 0.1, the batch size is set to 4. The total epoch of the training step is set to 100, which ensures the model is fully trained and the loss is converged. We leverage the pre-trained RoBERTa model. Besides, we also conducted a series of experiments to find the optimal event and relation weight ratio. We set the weight between event and TempRel from 1:1 to 20:1, and find that the model gets the best F1 score when the ratio is set to 10:1.

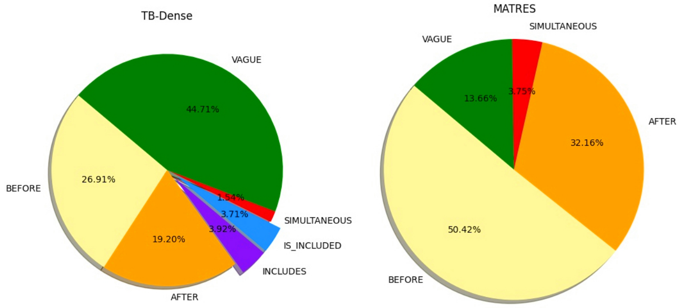


Fig. 3. Statistical analysis of labels on test sets.

3. EXPERIMENTS

In this section, we first provide experimental settings including baseline models and datasets. Then, we present the performances of our model in details, including overall performance, the evaluation on different TempRels, and ablation studies.

3.1. Baselines

We first use several feature-based systems CAEVO[3][4], CogCompTime[12][5] and Perceptron[13] as our baseline system. CAEVO adopts the Maximum Entropy(MaxEnt) method, CogCompTime uses the averaged perceptron method, and Perceptron proposes a multi-axis modeling to capture the events temporal structure. Conceptually, these methods mainly used manually designed features to train the model as a joint extraction system. Besides, we also employ the RNN+CSE+ILP model [2] and the BiLSTM+MAP model [4] as our baseline systems.

3.2. Datasets

We evaluate the performance of our model on two standard datasets: TimeBank-Dense [14] and MATRES [15]. They all use the TimeML-based[6] annotation schema between events[16]. We briefly analyzed the data sets and summarize the statistics of TB-Dense and MATRES as follows:

Data Info	#of Documents			#of Pairs		
	Train	Dev	Test	Train	Dev	Test
TB-Dense	22	5	9	4032	629	1427
MATRES	183	-	20	6332	-	827

Table 1. Dataset overview of TB-Dense and MATRES.

[4] https://www.usna.edu/Users/cs/nchamber/caevo/
[5] https://cogcomp.seas.upenn.edu/page/publication_view/844
[6] http://www.timeml.org/

3.3. Overall Performance

We summarize the performances of our proposed model and all baseline models in Table 2. We can see that our model achieved the state-of-the-art result in most cases. Specifically, our model outperforms CAEVO[3] by 0.7% in F1 on event extraction and 8.6% in F1 on TempRel extraction, and improves over CogCompTime[12] by 1.3% in F1 on event extraction and by 10.7% in F1 on TempRel extraction. Compared with Perceptron[13] and RNN+CSE+ILP[2], our model has 7.6% and 0.3% improvement on MATRES dataset respectively. Compared with BiLSTM+MAP[4], the performances of our model are different on the two datasets[7].For the TB-Dense dataset, our model improves over it by nearly 1.1% in F1 on TempRel extraction but drops by 1.1% in F1 on event extraction. For the MATRES dataset, our model improves over it by 0.1% in F1 on event extraction and by 1.1% in F1 on TempRel extraction.

Model(F1%)	TB-Dense		MATRES	
	Event	Relation	Event	Relation
CAEVO[3]	87.4	57.0	-	-
CogCompTime[12]	-	-	85.2	65.9
Perceptron[13]	-	-	-	69.0
RNN+CSE+ILP[2]	-	-	-	76.3
BiLSTM+MAP[4]	**89.2**	64.5	86.4	75.5
Our Model	88.1	**65.6**	**86.5**	**76.6**

Table 2. Model experiment results on TB-Dense and MATRES,both columns are reported in the previous papers.Bold indicate the best scores.

3.4. Single TempRel Performance

From Fig.2 we can see that BEFORE, AFTER labels occupy the main parts of the two datasets, and that INCLUDE, IS_INCLUDED, SIMULTANEOUS labels are very few. Such imbalanced distribution of TempRels make few-shot TempRels very hard to identify.

From Table 3, we can see that our model outperforms BiLSTM+MAP on BEFORE, AFTER, VAGUE and INCLUDE relations, and have similar performance on IS_INCLUDED and SIMULTANEOUS relations. We observe that the improvement on BEFORE, AFTER and VAGUE relation mainly benefit from introducing the external knowledge and using selective attention, which improve the precision of BEFORE/AFTER relation recognition and consequently increase the recall of VAGUE relation. In addition, the improvement on INCLUDES relation mainly benefit from data augmentation and Focal Loss components. But for the IS_INCLUDED label, two models have similar unsatisfied performances. For SIMULTANEOUS label, our model does not produce any correct prediction, which implies that SIMULTANEOUS relation is very hard to recognize. We will focus on this problem in the future work.

Model	BiLSTM+MAP			Our Model		
	P.	R.	F1.	P.	R.	F1.
B	70.4	58	63.6	**79.9**	58	**67.2**
A	61.0	69	64.7	**70.1**	66.7	**68.4**
I	23.8	9.0	13.0	**38.9**	12.5	**18.9**
II	40.0	30.1	34.4	46.7	26.4	33.7
S	-	-	-	-	-	-
V	61.2	67.2	64.1	61.4	**78.4**	**68.9**

Table 3. Model performance breakdown on TB-Dense dataset, Label abbreviations are explained below BEFORE(B),AFTER(A),INCLUDES(I),IS_INCLUDED(II), SIMULTANEOUS(S),VAGUE(V).

3.5. Ablation Tests

To further understand our proposed model, we conduct an ablation test by evaluating three main components, which illustrates the importance of each component of our framework.

Method(F1%)	TB-Dense	MATRES
All Components	65.6	76.6
w/o $Knowledge + Attention$	62.0	74.6
w/o $Attention$	62.3	75.1
w/o $DataAugmentation$	62.6	75.3
w/o $FocalLoss$	62.8	75.2

Table 4. Ablation Tests.

According to the Table 4, we have the following observations: For TB-Dense dataset, without any one of the three components, the performance of our model will drop about 3% in F1 scores. For MATRES dataset, without any one of the three components, the performance of our model will drop from 1% to 2% in F1 scores. Notably, if we only introduce external knowledge without selective attention, the performance of our model will drop dramatically. This illustrates the importance of each component in our framework.

4. CONCLUSION

In this paper we propose a novel neural model for joint event and temporal relation extraction, which integrates temporal commonsense knowledge, data augmentation and label balance loss function into one single system. The experimental results on two benchmark datasets show the effectiveness of our method and all three components.

[7] We download and run the code of BiLSTM+MAP[4] using the hyperparameters provided by its authors, and get the F1 values 87.8%, 61.5%, 86.1%, 74.3% in the same order in table 2. Here we presents the original F1 values in their paper

5. REFERENCES

[1] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky, "SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations," in *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June 2013, pp. 1–9, Association for Computational Linguistics.

[2] Qiang Ning, Sanjay Subramanian, and Dan Roth, "An improved neural baseline for temporal relation extraction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 6203–6209, Association for Computational Linguistics.

[3] N. Chambers, T. Cassidy, B. Mcdowell, and S. Bethard, "Dense event ordering with a multi-pass architecture," *Transactions of the Association for Computational Linguistics*, vol. 2, no. 11, pp. 273–284, 2014.

[4] Rujun Han, Qiang Ning, and Nanyun Peng, "Joint event and temporal relation extraction with shared representations and structured prediction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 434–444, Association for Computational Linguistics.

[5] Rujun Han, Yichao Zhou, and Nanyun Peng, "Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 5717–5729, Association for Computational Linguistics.

[6] Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer, "Richer event description: Integrating event coreference with temporal, causal and bridging annotation," in *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, Austin, Texas, Nov. 2016, pp. 47–56, Association for Computational Linguistics.

[7] J. Pustejovsky, P. Hanks, R Saurí, A. See, and Marcia Lazo, "The timebank corpus," *proceedings of corpus linguistics*, 2003.

[8] Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth, "Improving temporal relation extraction with a globally acquired statistical resource," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, June 2018, pp. 841–851, Association for Computational Linguistics.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] Y. Meng, A. Rumshisky, and A. Romanov, "Temporal information extraction for question answering using syntactic dependencies in an lstm-based architecture," 2017.

[11] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Q. Ning, B. Zhou, Z. Feng, H. Peng, and R. Dan, "Cogcomptime: A tool for understanding time in natural language," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018.

[13] Qiang Ning, Hao Wu, and Dan Roth, "A multi-axis annotation scheme for event temporal relations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 1318–1328, Association for Computational Linguistics.

[14] Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard, "An annotation framework for dense event ordering," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, June 2014, pp. 501–506, Association for Computational Linguistics.

[15] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth, "Joint reasoning for temporal and causal relations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 2278–2288, Association for Computational Linguistics.

[16] J. Pustejovsky, R. Ingria, R. Saur, J. Castano, and G. Katz, "The specification language timeml," *computational linguistics*, 2005.