Improving Event Temporal Relation Classification via Auxiliary Label-Aware Contrastive Learning

Tiesen Sun and Lishuang Li(✉)

School of Computer Science and Technology, Dalian University of Technology, Dalian, China
lils@dlut.edu.cn

Abstract. Event Temporal Relation Classification (ETRC) is crucial to natural language understanding. In recent years, the mainstream ETRC methods may not take advantage of lots of semantic information contained in golden temporal relation labels, which is lost by the discrete one-hot labels. To alleviate the loss of semantic information, we propose learning Temporal semantic information of the golden labels by Auxiliary Contrastive Learning (TempACL). Different from traditional contrastive learning methods, which further train the PreTrained Language Model (PTLM) with unsupervised settings before fine-tuning on target tasks, we design a supervised contrastive learning framework and make three improvements. Firstly, we design a new data augmentation method that generates augmentation data via matching templates established by us with golden labels. Secondly, we propose patient contrastive learning and design three patient strategies. Thirdly we design a label-aware contrastive learning loss function. Extensive experimental results show that our TempACL effectively adapts contrastive learning to supervised learning tasks which remain a challenge in practice. TempACL achieves new state-of-the-art results on TB-Dense and MATRES and outperforms the baseline model with up to 5.37%F_1 on TB-Dense and 1.81%F_1 on MATRES.

Keywords: Temporal relation classification · Contrastive learning

1 Introduction

The temporal relations of events are used to describe the occurring sequence of events in an article. Therefore understanding the temporal relations of events in articles is useful for many downstream tasks such as timeline creation [12], generating stories [4], forecasting social events [10], and reading comprehension [15]. Hence, the ETRC task is an important and popular natural language understanding research topic among NLP community.

The ETRC task is to determine the occurrence sequence of a given event pair. The context of the event pair is usually given to aid judgment. Ning et al. [14] first encoded the event pairs into embedded representations and then used

M. Sun et al. (Eds.): CCL 2022, LNAI 13603, pp. 180–193, 2022.
https://doi.org/10.1007/978-3-031-18315-7_12

fully connected layers as a classifier to generate confidence scores for each category of temporal relations. All related works of the NLP community since then have followed the classification view: classifying the embedded representations. Naturally, we can encode the context and events into a better embedding space in which the different relations are distinguished well, to get better classification results.

Traditionally, all recent works use one-hot vectors to represent golden temporal relation labels in the training stage. However, the one-hot vector reduces the label with practical semantics to the zero-one vector. It makes the embedded representations extracted by the ETRC models waiting for classifying be the similarities of the instances with the same label. But, the similarities are not equal to the label semantics, and lead to arbitrary prediction and poor model generalization, especially for confused instances. In brief, the one-hot vectors which represent temporal relation categories lose much semantic information.

To cope with the loss of semantic information in golden labels, we propose to learn the lost semantic information by contrastive learning, which is well confirmed and most competitive method for learning representations under unsupervised settings, so that the ETRC model can obtain better event representations. However, effectively adapting contrastive learning to supervised learning tasks remains a challenge in practice. General methods such as [3], which continue to train the PTLM model using unsupervised contrastive learning on the input texts (without labels) from the target task before fine-tuning, apply contrastive learning to supervised representation learning mechanically. They discard the category information in the process of further training. In the supervised ETRC task, we want the event pair representations with the same category to be as close as possible without collapsing. But direct application of the unsupervised contrastive learning loss function would prevent them from getting closer, because it discard the category information. It's an inherent problem of self-supervised contrastive learning. So the standard contrastive learning is not natural for the supervised ETRC task. To solve this problem we designed label-aware contrastive learning loss and design a new contrastive learning framework. Additionally, we argue that we can do contrastive learning in the intermediate layers of the PTLM as same as the last layer simultaneously. In a cascade structure, a change in previous layers affects the subsequent layers and continuous positive changes will make the learning process easier. Hence, we propose patient contrastive learning and design three patient strategies.

Overall, we propose TempACL: Firstly, we manually construct templates based on the semantics of labels and get augmentation sentences by matching the labels of instances. Secondly, we train the encoder of key samples which are necessary for contrastive learning by the augmentation datasets established by the ETRC datasets and the augmentation sentences. Thirdly, we jointly train the ETRC model with cross entropy loss and label-aware contrastive learning loss using a patient contrastive learning strategy.

The main contributions of this paper can be summarized as follows:

1. We propose learning the lost semantic information in golden labels by contrastive learning, and then design TempACL, a supervised contrastive learning framework based on a new data augmentation method designed by us. To our knowledge, we are the first to propose using contrastive learning on the ETRC task.
2. In order to make our TempACL achieve better performance, we design label-aware contrastive learning loss and patient contrastive learning strategy.
3. We demonstrate the effectiveness of our TempACL on TB-Dense and MATRES datasets. Our TempACL outperforms the current best models with up to 2.13%F_1 on TB-Dense and 1.26%F_1 on MATRES and outperforms the baseline model with up to 5.37%F_1 on TB-Dense and 1.81%F_1 on MATRES.

2 Related Work

2.1 Event Temporal Relation Classification

Since the birth of pre-trained language models, researchers have mainly used them to encode event representations and design many new methods based on them. Wang et al. [19] propose a JCL method that makes the classification model learn their designed logical constraints within and across multiple temporal and subevent relations by converting these constraints into differentiable learning objectives. Zhou et al. [24] propose the CTRL-PG method, which leverages the Probabilistic Soft Logic rules to model the temporal dependencies as a regularization term to jointly learn a relation classification model. Han et al. [8] propose the ECONET system, which further trains the PTLM with a self-supervised learning strategy with mask prediction and a large-scale temporal relation corpus. Zhang et al. [23] propose the TGT network that integrates both traditional multi-head self-attention and a new temporal-oriented attention mechanism and utilizes a syntactic graph that can explicitly find the connection between two events. Tan et al. [18] propose the Poincaré Event Embeddings method which encodes events into hyperbolic spaces. They argue that the embeddings in the hyperbolic space can capture richer asymmetric temporal relations than the embeddings in the Euclidean space. And they also proposed the HGRU method which additionally uses an end-to-end architecture composed of hyperbolic neural units, and introduces common sense knowledge [14].

All of the above methods use the one-hot vector and lose the semantic information of the golden label. To take advantage of the missing semantic information, we make the target ETRC model learn from them via contrastive learning.

2.2 Contrastvie Learning

Contrastive learning aims to learn efficient representations by pulling semantically close neighbors together and pushing non-neighbors away [7]. In recent years, self-supervised contrastive learning and supervised contrastive learning have attracted more and more researchers to study them.

Self-supervised Contrastvie Learning. In computer vision (CV), We et al. [21] propose MemoryBank, which maintain a large number of representations of negative samples during training and update negative sample representations without increasing batch size. He et al. [9] propose MoCo, which designs the momentum contrast learning with two encoders and employs a queue to save the recently encoded batches as negative samples. Chen et al. [2] proposed the SimCLR which learns representations for visual inputs by maximizing agreement between differently augmented views of the same sample via a contrastive loss. Grill et al. [5] propose BYOL, which uses asymmetric two networks and discards negative sampling in self-supervised learning. In Natural Language Processing (NLP), Yan et al. [22] propose ConSERT, which has a similar model structure to SimCLR, except that ResNet is replaced by Bert and the mapping header is removed. And they also propose multiple data augmentation strategies for contrastive learning, including adversarial attack, token shuffling, cutoff and dropout.

Supervised Contrastvie Learning. Khosla et al. [11] extend the self-supervised contrastive approach to the fully-supervised setting in the CV domain, and take many positives per anchor in addition to many negatives (as opposed to self-supervised contrastive learning which uses only a single positive). Gunel et al. [6] extends supervised contrastive learning to the NLP domain with PTLMs.

Different from ConSERT we design a new data augmentation method based on templates in our contrastive learning framework. And different from Khosla's work, we design a new supervised contrastive loss which still uses only a single positive but does not treat the sentence representations with the same label as negative examples.

3 Our Baseline Model

Our baseline model is comprised of an encoder and a classifier. We use RoBERTa [13] as our encoder and use two fully connected layers and a tanh activation function between them as our classifier. Recently, most of the related works use RoBERTa as an encoder, because RoBERTa can achieve better results on the ETRC task than BERT in practice.

Each instance is composed of an event temporal triplet t (i.e. ($< e_1 >$, $< e_2 >$, r), where $< e_1 >$ and $< e_2 >$ are event mentions and r is the temporal relation of the event pair.) and the context s of the events which may be a single sentence or two sentences.

We first tokenize the context and get a sequence of tokens $X_{[0,n)}$ with length n. Then we feed the $X_{[0,n)}$ into RoBERTa. One event mention may correspond to multiple tokens, so we send the token embeddings corresponding to these tokens to an average pooling layer to get the final event representation e_i. Next, we combine e_1 and e_2 into a classification vector $e_1 \oplus e_2$, where $\oplus$ is used to denote concatenation. Finally, we feed the classification vector into the classifier

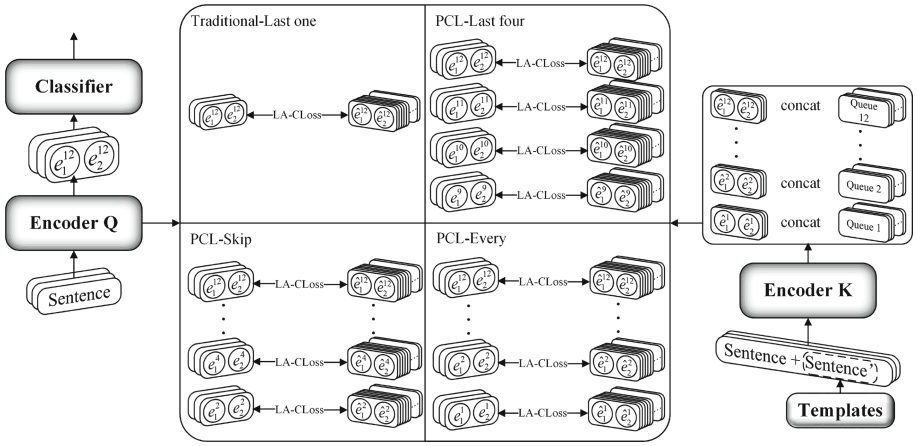

Fig. 1. Joint training with patient contrastive learning. We name the PLTM which encodes positive and negative key samples as Encoder K and the PLTM used for ETRC as Encoder Q.

followed by a soft-max function to get confidence scores for each category of temporal relations.

4 Self-supervised Contrastive Learning

Contrastive learning is learning by pulling similar instance pairs closer and pushing dissimilar instance pairs farther. The core of self-supervised contrastive learning is to generate augmented examples of original data examples, create a predictive task where the goal is to predict whether two augmented examples are from the same original data example or not, and learn the representation network by solving this task. He et al. [9] formulate contrastive learning as a dictionary look-up problem and propose an effective contrastive loss function L_{CL} with similarity measured by dot product:

$$L_{CL} = -\log \frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{\{K^{-}\}} \exp\left(q \cdot k^{-}/\tau\right)} \tag{1}$$

where q is a query representation, k^{+} is a representation of the positive (similar) key sample, k^{-} are representations of the negative (dissimilar) key samples, K^{-} is a negative key samples set, and τ is a temperature hyper-parameter. He et al. [9] also propose maintaining the dictionary as a queue of data samples. It allows contrastive learning to reuse the previous batch of key samples so that we can increase the number of negative samples without increasing the batch size, thus improving the performance of the model. The dictionary size is a flexible hyper-parameter. The samples in the dictionary are progressively replaced. The current batch is enqueued to the dictionary, and the oldest batch in the queue

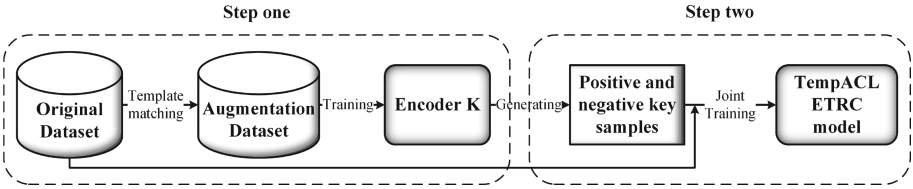

Fig. 2. Overall process of TempACL

Table 1. Templates. All the six temporal relation labels are in TB-Dense and * indicates the temporal relation label also exists in MATRES.

Temporal relation	Templates
AFTER*	the beginning of the event of $< e_1 >$ is after the end of the event of $< e_2 >$
BEFORE*	the end of the event of $< e_1 >$ is before the beginning of the event of $< e_2 >$
INCLUDES	the beginning of the event of $< e_1 >$ is before the beginning of the event of $< e_2 >$ and the end of event of $< e_1 >$ is after the end of the event of $< e_2 >$
IS_INCLUDED	the beginning of the event of $< e_1 >$ is after the beginning of the event of $< e_2 >$ and the end of event of $< e_1 >$ is before the end of the event of $< e_2 >$
VAGUE*	the temporal relation between the event of $< e_1 >$ and the event of $< e_2 >$ is vague
SIMULTANEOUS*	the event of $< e_1 >$ and the event of $< e_2 >$ have the same beginning and end time

is removed. In this paper, we follow this part of their work and transfer it to the supervised ETRC task.

5 TempACL Approach

In this section, we introduce our TempACL approach in details and draw the overall process of TempACL in Fig. 2. TempACL aims to encoder semantic information of golden temporal relation labels and uses contrastive learning to make the baseline model extract better event representations. Hence, we first train Encoder K used for encoding semantic information of golden temporal relation labels, and then jointly train the baseline model with auxiliary contrastive learning via the label-aware contrastive learning loss function and a patient strategy. Specially, we fix the parameters of the Encoder K in the joint training stage.

5.1 Training Encoder K

First of all, we need to establish templates. In order to make the positive key samples encoded by Encoder K contain as much and as detailed semantic infor-

mation of golden temporal relation labels as possible, we need to create efficient templates that automatically convert each golden temporal relation label into a temporal information-enriched sentence s' to enrich the semantic information of golden temporal relation labels. We argue that the time span of events (i.e., the duration of the events) guides ETRC. So we use the start and end times of events and the temporal relation between events to describe the temporal relation of the event pair on a subtle level. We show the templates in Table 1.

Subsequently, we build the augmentation dataset. For each record (t, s) in original Dataset, we use r to match the templates and get s' by filling events into the corresponding positions in the template, then concatenate s and s' to get an augmentation sentence $s_{aug} = s + s'$, finally get a new record (t, s_{aug}). We combine all new records into an augmentation dataset.

Finally, we use the augmentation dataset to train the Encoder K with the help of the classifier which we propose in Sect. 3 under supervised setting. Encoder K is a RoBERTa model.

5.2 Joint Training with Patient Label-Aware Contrastive Loss

The trained Encoder K has been obtained, we can start joint training in Fig. 1. We send s in the original dataset to Encoder Q, and then get event pair representations $\{e_{1j} \oplus e_{2j}\}_{j=1}^{12}$ in different layers of Encoder Q. e_{ij} is the hidden state corresponding to the event i from the j-th RoBERTa Layer. We simultaneously send s_{aug} in the augmentation dataset to Encoder K, and then get event pair representations $\{\hat{e}_{1j} \oplus \hat{e}_{2j}\}_{j=1}^{12}$ in different layers of Encoder K. $\hat{e}_{ij}$ is the hidden state corresponding to the event i from the j-th RoBERTa Layer, and $\hat{}$ is used to denote the hidden state from the Encoder K. We normalized $e_{1j} \oplus e_{2j}$ as the query q and $\hat{e}_{1j} \oplus \hat{e}_{2j}$ as key k with L2 Norm. According to different patient strategies, queries and keys of different layers were selected for comparative learning.

We should not mechanically apply the loss function of self-supervised contrastive learning in Eq. 1 to the supervised ETRC directly. In the supervised ETRC task, we want the event pair representations with the same category to be as close as possible without collapsing. But L_{CL} treat the key samples in the queue, whose event pair have the same temporal relation with the event pair of the query sample, as negative key samples. Therefore, in the process of minimizing the L_{CL}, the event pair representations with the same category are mutually exclusive, which confuse the ETRC model. So we propose label-aware contrastive loss function L_{LACL}:

$$L_{LACL} = -\sum_{i=1}^{N} \left(\log \frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{\{K'^{-}\}} \exp\left(q \cdot k'^{-}/\tau\right)} \right)_i \tag{2}$$

where $\bar{K}^{-}$ is negative key samples set which except the key samples with the same label as q, and N is the number of training samples. In practice, we convert $q \cdot k$ where $k \in \{k : k \in K^{-}, k \notin K'^{-}\}$ to -10^6 by matrix operations.

Inspired by Sun et al. [17], we argue that using the event pair representations of the intermediate layers of the Encoder Q and the event pair representations of the intermediate layers of the Encoder Q for additional contrastive learning can enhance the learning of semantics of the Encoder Q, and improve the performance of the baseline model. Hence we propose patient label-aware contrastive learning loss L_{PCL} based on Eq. 2:

$$L_{PCL} = -\sum_{j \in J} \sum_{i=1}^{N} \frac{1}{\|J\|} \left(\log \frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{\{K'^{-}\}} \exp\left(q \cdot k'^{-}/\tau\right)} \right)_{i,j} \tag{3}$$

where J is the set of intermediate layers involved in contrastive learning. Specifically, we propose three patient contrastive learning strategies: (1) PCL-Last four: we contrast the last four layers of the Encoder Q and Encoder K (Fig. 1 upper right). (2) PCL-Skip: we contrast every two layers of the Encoder Q and Encoder K (Fig. 1 lower left). (3) PCL-Every: we contrast every layers of the Encoder Q and Encoder K (Fig. 1 lower right).

Finally, we jointly train ETRC task and auxiliary label-aware contrastive learning task with the final loss function L_final:

$$L_{finall} = \alpha L_{CE} + \beta L_{PCL} \tag{4}$$

where L_{CE} is cross-entropy loss function, α and β are hyper-parameters which weight the importances of ETRC task and auxiliary label-aware contrastive learning task.

6 Experiments and Results

In this section, we perform experiments on TB-Dense and MATERS and prove our TempACL performs better than previous state-of-the-art methods. Details on the datasets, experimental setup, and experimental results are provided in the following subsections.

TB-Dense TB-Dense [1] is a densely annotated dataset for the ETRC and annotated based on TimeBank. It also annotates the temporal relations of pairs of events across sentences, different from TimeBank which only annotates events in the same sentence. It annotates a total of 6 temporal relations (AFTER, BEFORE, INCLUDE, IS INCLUDED, VAGUE, SIMULTANEOUS). We follow the split strategy of Han et al. [8] and Zhange et al. [23] which uses 22 documents as train set, 5 documents as dev set and 9 documents as test set.

MATERS MATERS [16] is refined from 275 documents in TimeBank and TempEval (containing AQUAINT and Platinum). Ning et al. [16] design a novel multi-axis (i.e., main, intention, opinion and hypothetical axes) annotation scheme to further annotate the 275 documents. There are only 4 temporal

Table 2. Data statistics for TB-Dense and MATRES

	TB-Dense		MATRES	
	Documents	Triplets	Documents	Triplets
Train	22	4032	204	10097
Dev	5	629	51	2643
Test	9	1427	20	837

relations (BEFORE, AFTER, EQUAL and VAGUE) different from TB-Dense and the EQUAL is the same as SIMULTANEOUS. We follow the official split strategy that uses TimeBank and AQUAINT for training and Platinum for testing. We also follow the previous works [14,18] that randomly select 20 percents of the official train documents as dev set.

We briefly summarize the data statistics for TB-Dense and MATRES in Table 2.

6.1 Dataset

6.2 Experimental Setup

In the process of training Encoder K, we add a dropout layer between the Encoder K and the Classifier and set the drop probability to 0.5, in order to make the key samples contain more useful temporal information. We train Encoder K 10 and 20 epochs respectively on TB-Dense and MATRES. We set the batch size to 24, the τ to 0.1, the learning rate of the Classifier to 5e−4 and the learning rate of RoBERTa to 5e−6. We use grid search strategy to select the best $\alpha \in [0.7: 1.4]$ and $\beta \in [0.01: 0.001]$. As for the dimension of the hidden states between two fully connected layers in the Classifier, we set it to 36. We set the size of the queue to 3840 and 9600 respectively on TB-Dense and MATRES.

6.3 Main Results

As shown in Table 3, we compare our approach with other state-of-the-art methods in recent years on TB-Dense and MATRES. We report the best F_1 value for each method. The compared methods have been introduced in Sect. 2. And the results of compared methods are directly taken from the cited papers except CERT[1]. We reproduce CERT and record the results.

We observe that our baseline model achieves 63.56%F_1 on TB-Dense and 79.95%F_1 on MATRES. It demonstrates that our baseline model can effectively classify temporal relation, and even achieves a competitive performance that is close to the current best 80.5%F_1 on MATRES. Furthermore, our TempACL outperforms previous state-of-the-art methods on ETRC with up to 2.13%F_1 on TB-Dense and 1.26%F_1 on MATRES. Compared with CERT, the traditional

[1] https://github.com/UCSD-AI4H/CERT.

Table 3. Comparison of various approaches on ETRC on TB-Dense and MATRES. Bold denotes the best performing model. F_1-score (%)

Method		TB-Dense	MATRES
JCL [19]	RoBERTa base	-	78.8
ECONET [8]	RoBERTa Large	66.8	79.3
TGT [23]	BERT Large	66.7	80.3
Poincaré event embeddings [18]	RoBERTa base	–	78.9
HGRU+knowledge [18]	RoBERTa base	–	80.5
CERT [3]	RoBERTa base	64.92	80.46
Baseline (ours)	RoBERTa base	63.56	79.95
TempACL (ours)	RoBERTa base	**68.93**	**81.76**

self-supervised contrastive learning method, our TempACL achieves 4.01%F_1 and 1.30%F_1 improvement respectively. These experimental results prove the effectiveness of learning semantic information of golden temporal relation labels via patient label-aware contrastive learning. There are three possible reasons for the effectiveness: (1) The difference between the query representation and the key representation comes from the semantic information of the golden temporal relation label, because the input of Encoder Q doesn't have the label information but the input of Encoder K input does. The L_{LACL} forces q closer to K to reduce the difference. So that in the process of minimizing L_{LACL} Encoder Q learns the label semantic information and forces itself to extract more useful information related to golden temporal relation labels from the sentences that do not contain any golden temporal relation label information. (2) The supervised contrastive learning framework and L_{LACL} designed by us is more suitable for the ETRC task than the traditional self-supervised contrastive learning method. (3) The data augmentation method proposed by us not only utilizes the semantic information of labels but also enriches the semantic information of labels.

Different from JCL and HGRU, which use external commonsense knowledge to enrich the information contained in event representations, TempACL enables the model to better mine the information contained in original sentences. Compared to ECONET and TGT, which use a larger pre-trained language model, or TGT and HGRU, which use networks with complex structures followed RoBERTa base or BERT Large, TempACL enables a smaller and simpler model which only contains a RoBERTa base and two fully connected layers to achieve the state-of-the-art performance.

6.4 Ablation Study and Qualitative Analysis

We observe that, TempACL make improvements of 5.37%F_1 and 1.81%F_1 on TB-Dense and MATRES respectively compared with the baseline model. In this section, we first qualitatively analyze key samples, and then we do the ablation experiments to further study the effects of patient strategies and label-aware

Table 4. Results of TempACL with different strategies. F_1-score (%)

Method	TB-Dense	MATRES
Traditional-last one	66.17	80.95
PCL-Last four	68.93	81.76
PCL-Skip	67.73	80.46
PCL-Every	65.23	80.37

Table 5. Results of TempACL with different contrastive learning loss. F_1-score (%)

Method	TB-Dense	MATRES
TempACL-LACL	68.93	81.76
TempACL-TCL	66.03	80.89
Baseline	63.56	79.95

contrastive learning loss. We ensure that all ablation results are optimal by using optimal strategies under the given conditions.

Qualitative Analysis. Wang et al. [20] propose to justify the effectiveness of contrastive learning in terms of simultaneously achieving both alignment and uniformity. Hence we reduce the dimension of key samples in each layer through PCA and represent it in Fig. 3 on TB-Dense. All four contrastive strategies we used to utilize the key samples of the last layer, so we take Fig. 3(l) to analyze the alignment and uniformity of TempACL. On the one hand, we can see that there are 6 clusters of representations that are well-differentiated even in two dimensions. Our method maps key samples with the same category to a relatively dense region. These well demonstrate that our embedded knowledge has a strong alignment. On the other hand, we also can see that the 5 clusters, which represent temporal categories in Fig. 3(l) right, are farther from the VAGUE cluster than each other. It means that our embedded knowledge retains as much category information as possible. The farther away different clusters are, the more category information and differences are retained. Moreover, different key samples with the same category distribute evenly within the dense region, which means that our key samples retain as much instance information as possible. Furthermore, the more evenly distributed they are, the more information they retain. These well demonstrate that our embedded knowledge has a strong uniformity. We find that the key samples encoded by the last four layers of the Encoder K have strong alignment and uniformity.

Last One Strategy vs Patient Strategy. In Sect. 5.2 we propose three patient strategies. In this section, we do experiments to study which strategy is optimal and report the experimental results in Table 4. PCL-Last four achieves the best results on both TB-Dense and MATRES. On the one hand, PCL-Last

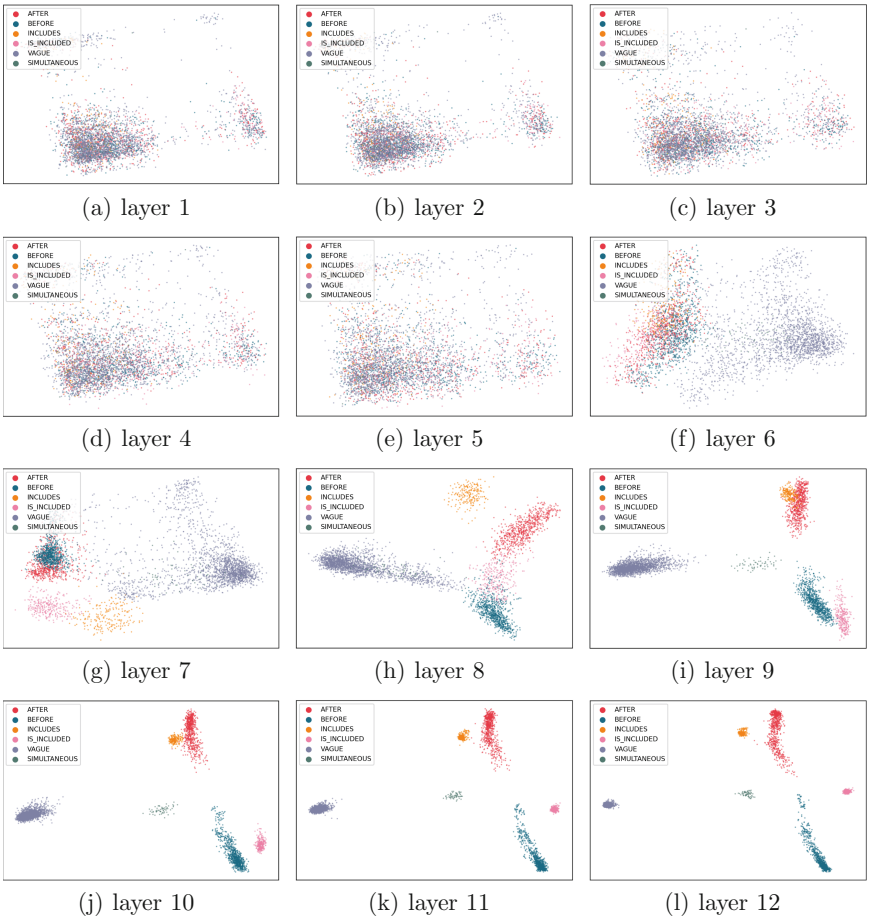

Fig. 3. The distributions of key samples of each RoBERTa layers on TB-Dense.

four provides more positive and negative samples. In Fig. 3, the distribution of key samples in the last four layers also indicates that these positive and negative samples have great value in learning. On the other hand, this layer-by-layer approach greatly reduces the difficulty of learning. In the PTLM, different sublayers are cascade, and the changes in the output in the front layers influence the latter layers. PCL-every performs poorly and worse than Traditional-Last one, because the first eight layers do not provide good positive and negative key samples, and learning them confuses the model. However PCL-Skip performs better than Traditional-Last one. This is because the number of bad key samples in PCL-Skip is relatively small, which makes the negative impact of these bad key samples much smaller. The layer-by-layer approach reduces the difficulty of learning and the benefits outweigh the negative impact.

Label-Aware Contrastive Loss vs Traditional Contrastive Loss. In order to determine whether our proposed label-aware contrastive loss has a positive effect, we conduct a comparative experiment and record the experimental results in Table 5. We compare the TempACL with label-aware contrastive learning loss (TempACL-LACL) and the TempACL with traditional contrastive learning loss (TempACL-TCL) on TB-Dense and MATRES respectively. We can see that the TempACL-LACL achieves 2.90%F_1 and 0.87%F_1 performance improvement over the TempACL-TCL respectively. It shows the benefit of eliminating key samples with the same label as the query from the negative samples set. The reason is that using key samples, which have the same label as the query, as negative samples prevent instances of the same label from learning similar event representations to some extent, which runs counter to the ETRC's aims. And the label-aware contrastive learning loss can avoid such a situation.

7 Conclusion

In recent years, the mainstream ETRC methods focus on using discrete values to represent temporal relation categories and lose too much semantic information contained in golden labels. So we propose TempACL, which makes the ETRC model learn the lost semantic information in golden labels via contrastive learning. Extensive experiments prove the contrastive learning framework in TempACL is more suitable for the supervised ETRC task than traditional self-supervised contrastive learning. The patient contrastive learning strategy designed by us provides more useful positive and negative key samples and reduces the difficulty of contrastive learning. The label-aware contrastive learning loss designed by us avoids the negative interactions between different queries and keys in the same category, which is an inherent problem of self-supervised contrastive learning.

Acknowledgments. This work is supported by grant from the National Natural Science Foundation of China (No. 62076048), the Science and Technology Innovation Foundation of Dalian (2020JJ26GX035).

References

1. Cassidy, T., McDowell, B., Chambers, N., Bethard, S.: An annotation framework for dense event ordering. In: ACL (Volume 2: Short Papers), pp. 501–506 (2014)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp. 1597–1607. PMLR (2020)
3. Fang, H., Wang, S., Zhou, M., Ding, J., Xie, P.: Cert: contrastive self-supervised learning for language understanding. arXiv e-prints pp. arXiv-2005 (2020)
4. Goldfarb-Tarrant, S., Chakrabarty, T., Weischedel, R., Peng, N.: Content planning for neural story generation with aristotelian rescoring. In: EMNLP, pp. 4319–4338 (2020)

5. Grill, J.B., et al.: Bootstrap your own latent - a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 21271–21284 (2020)
6. Gunel, B., Du, J., Conneau, A., Stoyanov, V.: Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403 (2020)
7. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2, pp. 1735–1742 (2006)
8. Han, R., Ren, X., Peng, N.: ECONET: effective continual pretraining of language models for event temporal reasoning. In: EMNLP, pp. 5367–5380 (2021)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9726–9735. IEEE (2020)
10. Jin, W., et al.: Forecastqa: a question answering challenge for event forecasting with temporal text data, pp. 4636–4650 (2021)
11. Khosla, P., et al.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)
12. Leeuwenberg, A., Moens, M.F.: Temporal information extraction by predicting relative time-lines. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)
13. Liu, Y., et al: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. Ning, Q., Subramanian, S., Roth, D.: An improved neural baseline for temporal relation extraction. In: EMNLP-IJCNLP, pp. 6203–6209 (2019)
15. Ning, Q., Wu, H., Han, R., Peng, N., Gardner, M., Roth, D.: TORQUE: a reading comprehension dataset of temporal ordering questions. In: EMNLP, pp. 1158–1172 (2020)
16. Ning, Q., Wu, H., Roth, D.: A multi-axis annotation scheme for event temporal relations. In: ACL (Volume 1: Long Papers), pp. 1318–1328 (2018)
17. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for BERT model compression. In: EMNLP-IJCNLP, pp. 4323–4332 (2019)
18. Tan, X., Pergola, G., He, Y.: Extracting event temporal relations via hyperbolic geometry. In: EMNLP, pp. 8065–8077 (2021)
19. Wang, H., Chen, M., Zhang, H., Roth, D.: Joint constrained learning for event-event relation extraction. In: EMNLP, pp. 696–706 (2020)
20. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning, pp. 9929–9939. PMLR (2020)
21. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination. arXiv preprint arXiv:1805.01978 (2018)
22. Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., Xu, W.: Consert: a contrastive framework for self-supervised sentence representation transfer. arXiv e-prints pp. arXiv-2105 (2021)
23. Zhang, S., Huang, L., Ning, Q.: Extracting temporal event relation with syntactic-guided temporal graph transformer. arXiv preprint arXiv:2104.09570 (2021)
24. Zhou, Y.,et al.: Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In: AAAI, vol. 35, pp. 14647–14655 (2021)