TEMPLATE: TempRel Classification Model Trained with Embedded Temporal Relation Knowledge

Anonymous ACL submission

Abstract

In recent years, the mainstream Temporal Relation (TempRel) classification methods may not take advantage of the large amount of semantic information contained in golden TempRel labels which is lost by the traditional discrete one-hot labels. So we propose a new approach that can make full use of golden TempRel label information and make the model performance better. Firstly we build a TempRel Classification model [1], which consists of a RoBERTa and a Classifier. Secondly we establish fine-grained templates to automatically generate sentences to enrich golden TempRel label information and build an Enhanced Data-set. Thirdly we use the Enhanced Data-set to train the Knowledge Encoder which has the same structure as the TempRel Classification model, and get embedded knowledge. Finally we Trian the TempRel Classification model with EMbedded temPoral reLATion knowldgE (TEMPLATE) by using our designed Cosine balanced MSE loss function. Extensive experimental results shows that our approach achieves new state-of-the-art results on TB-Dense and MATRES and outperforms the TempRel Classification model trained with only traditional cross entropy loss function with up to $5.51\%F_1$ on TB-Dense and $2.02\%F_1$ on MATRES. [2]

1 Introduction

Articles such as news usually describe a series of events with different start and end times. These events seem to be narrated discretely, but in fact there are certain connections. The most important type of event connection is the Temporal Relation (TempRel). It represents the sequence of events, which connects the development and evolution of the events in the article. If we can accurately extract the TempRel of events in the article, it will help many downstream tasks such as reading comprehension (Ning et al., 2020; Zhou et al., 2019), tracking biomedical histories (Sun et al., 2019; Bethard et al., 2016, 2017), generating stories (Yao et al., 2019; Goldfarb-Tarrant et al., 2020), and forecasting social events (Li et al., 2020; Jin et al., 2021). Therefore, the TempRel classification has always been an important natural language processing task, which has attracted more and more attention in the NLP community.

The TempRel classification task is to determine the relationship of the event pair in the candidate relationship set, given two events and one or two sentences of document (where each event is a span of the sentences). Naturally, all state-of-the-art models follow the classification view: the sentences and events are encoded as an embedded representation and then classified as one of the candidate relations. The training goal then aims to embed the sentences and events into a space in which the different relations are well separated. Recently the mainstream TempRel classification methods use pre-trained language models to encode event representations and concatenate them, then feed them into a classifier.

On the TempRel classification task, we can see that all state-of-the-art models (Ma et al., 2021; Wang et al., 2020; Han et al., 2021; Zhang et al., 2021; Tan et al., 2021; Ning et al., 2019) represent a golden TempRel label as a one-hot vector in training their classification models. But the discrete values which represent TempRel categories lost abundant semantic information. And one-hot labels assume that all categories are independent with each other. However in real situations, TempRel labels are not completely independent and have their own internal connections to each other which are well expressed by the semantic information contained in the golden TempRel labels. So if we can take advantage of this semantic information, the TempRel classification performance of the model will definitely be improved.

[1] Without special instruction, all the mentions of "TempRel Classification model" in the following are the model introduced in section 3

[2] We will release our code and data upon acceptance.

On the pre-trained language model distillation task, we can see that (Sun et al., 2019; Sanh et al., 2019; Jiao et al., 2020; Sun et al., 2020) compress pre-trained language model into a smaller one, such as compress BERT_{12} (the original BERT-Base (Devlin et al., 2019) model which has 12 layers of Transformer) into a smaller BERT_6 (same structure as BERT-Base but with only 6 layers of Transformer), meanwhile BERT_6 performs on-par with BERT_{12} (Jiao et al., 2020). BERT_6 has the same training data as BERT_{12} but fewer parameters. It indicates that general pre-trained language models have enough parameters and would have performed better on most downstream tasks. i.e. they have a lot of potential which is not exploited well. Their works inspire us that we can improve the TempRel classification performance of the model by using an additional loss functions such as mean square error loss to make the model learn from the embedded knowledge of golden TempRel labels.

Naturally we build a TempRel Classification model, which consists of a RoBERTa (Liu et al., 2019) and a Classifier. And we propose a new approach that can improve the performance of the model by learning embedded knowledge which consists of golden TempRel labels and templates which we build in section 4.1 with an additional loss function which we design in section 4.3.

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

1. In this paper, we build a TempRel Classification model and propose a new approach. We use templates to enrich the TempRel label information and encode it as embedded knowledge. Then we use the embedded knowledge and an additional loss function to make TempRel Classification model perform better.

2. In order to make our approach achieve better performance, we further design fine-grained templates and Cosine balanced MSE loss function as an additional loss function.

3. We demonstrate the effectiveness of our approach on TB-Dense and MATRES data-sets. Our approach outperforms the current best model with up to 2.27%F_1 on TB-Dense and 1.47%F_1 on MATRES and outperforms TempRel Classification model trained with traditional cross entropy loss function with up to 5.51%F_1 on TB-Dense and 2.02%F_1 on MATRES.

2 Related work

In this section, we introduce some works related to TempRel classification, label embedding and pre-trained language model distillation.

2.1 TempRel Classification

The TempRel classification task has always been a popular research topic among NLP community. (Cheng and Miyao, 2017) introduce the structure of Bi-directional Long Short-Term Memory (BI-LSTM) into their model, and encode the sentence sequence and the dependent sequence between the event nodes. Finally, the two codes are aggregated and used for classification. (Ning et al., 2019) also introduce Long Short-Term Memory (LSTM) network to encode the event features which take into account global contexts, and feed their representations into a multi-layer perception (MLP) for TempRel classification. They also introduce prior knowledge from the Temporal Common Sense Knowledge Base in the baseline model to assist classification. (Han et al., 2019) propose a combination of LSTM and Structured Support Vector Machine (SSVM). LSTM is used to learn the scoring function of the relational classifier, and SSVM replaces the traditional Integer Linear Programming process for global Maximum posterior inference. (Wang et al., 2020) propose a joint constrained learning framework which enforces logical constraints within and across multiple temporal and sub-event relations by converting these constraints into differentiate learning. (Zhou et al., 2021) propose Clinical Temporal ReLation Exaction with Probabilistic Soft Logic Regularization and Global Inference (CTRL-PG), 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., 2021) propose a further-training strategy, ECONET, to further train pre-trained language models with a large-scale temporal relationship corpus which makes pre-trained language models focus on the event time relationship in the sentence. Additionally ECONET performs better than the existing general language model under a low-resource setting through adequate experiments. (Zhang et al., 2021) propose a new Temporal Graph Transformer network and utilize syntactic graph which can explicitly find the connection between two events. (Tan et al., 2021) claim that the embedding in the Euclidean space cannot capture richer asymmetric temporal relations, therefore they embed events into hyperbolic spaces. Then they design a hyperbolic neural network and incorporate temporal commonsense. Unlike their methods, we enable a simple TempRel Classification model to achieve state-of-the-art results without using extra data by encoding golden TempRel label information and using an additional loss function.

2.2 Label Embedding

Label embedding is to learn the embedding of the labels in general domain classification tasks and has been proven to be effective. (Zhang et al., 2018) propose Multi-Task Label Embedding (MTLE) to convert labels in text classification task into semantic vectors and turn the original tasks into vector matching tasks. It uses input sequences from three or more tasks and learns along with their labels which benefits from each other and obtaining better sequence representations. (Wang et al., 2018) propose to regard text classification task as a label-word joint embedding problem: each label is embedded in the same space with the word vectors and introduce attention mechanism. (Yang et al., 2018) use label embedding and apply a sequence generation model to classify text which takes the correlations between labels into account. (Guo et al., 2021) propose a novel Label Confusion Model (LCM) which can learn label confusion to capture semantic overlap among labels by calculating the similarity between instance and labels during training and generate a better label distribution. In our work, we also encode label information, but in a different way. We use templates to convert labels into sentences which have more TempRel information, then use a new way to encode them.

2.3 Pre-trained Language Model Distillation

Pre-trained Language Model distillation has been proven a promising way to compress the large models while maintaining accuracy. (Tang et al., 2019) distill BERT-large into a single-layer Bi-LSTM, reducing the number of parameters by 100 times and increasing the speed training process by 15 times. Although the effect is much worse than BERT, it can be tied with ELMo. Different from previous studies, (Sun et al., 2019) propose Patient Knowledge Distillation, which extracts knowledge from the intermediate layers of the teacher model to avoid the phenomenon of over-fitting when only the last layer is distilled. The previous work is to distill the fine-tuned BERT, and the smaller model learns from task-related knowledge. (Sanh et al., 2019) propose a new strategy which performs distillation in the pre-training stage. Further, (Jiao et al., 2020) propose a two-stage learning framework, which distills the larger model in the pre-trained and fine-tuning stages respectively, and obtains excellent results. Inspired by these studies, we could make the intermediate layers of the model learn from the embedded knowledge in the fine-tuning stage.

3 Our Baseline Model

Our TempRel Classification model (i.e. baseline model), represented in Figure 1, is comprised of a pre-trained language model and a classifier which consists of two fully connected layers and a tanh activation function between them. We use RoBERTa (Liu et al., 2019), which has been proven to be more effective than BERT in TempRel classification task by (Zhao et al., 2021; Tan et al., 2021) for domain transfer, as our pre-trained language model.

For a given example (s, e_1, e_2, r), s is a span of document, which may be a single sentence or two sentences. e_1 and e_2 are events and each of them is represented as a span of s. r is a golden TempRel label from the label set {AFTER, BEFORE, INCLUDE, IS INCLUDED, VAGUE, SIMULTANEOUS}.

We first tokenize s to get a sequence $X_{[0,n)}$ of n tokens i.e. $\{x_0, x_1, \cdots, x_{n-1}\}$. Then we feed the sequence $X_{[0,n)}$ into RoBERTa and get the event contextual representations e_1^{12} and e_2^{12} corresponding to the tokens of e_1 and e_2 respectively. If the number of tokens of an event is larger than one, we use the mean value of the hidden states of all the tokens of the event as the event contextual representation. Next, we combine e_1^{12} and e_2^{12} into a classification vector $c = e_1^{12}; e_2^{12}$ where ; is used to denote concatenation. Finally we feed c into the classifier followed by a soft-max function to get a distribution over the six TempRel labels.

4 TEMPLATE Approach

Considering the missing information in discrete values which represent TempRel categories and the huge inherent underutilized ability of pre-trained language model, we aspire to propose a new approach to solve the above weaknesses together. To this end, we train TempRel Classification model with embedded knowledge of TempRel information and further focus on conquering two key questions. One is how to efficiently enrich TempRel

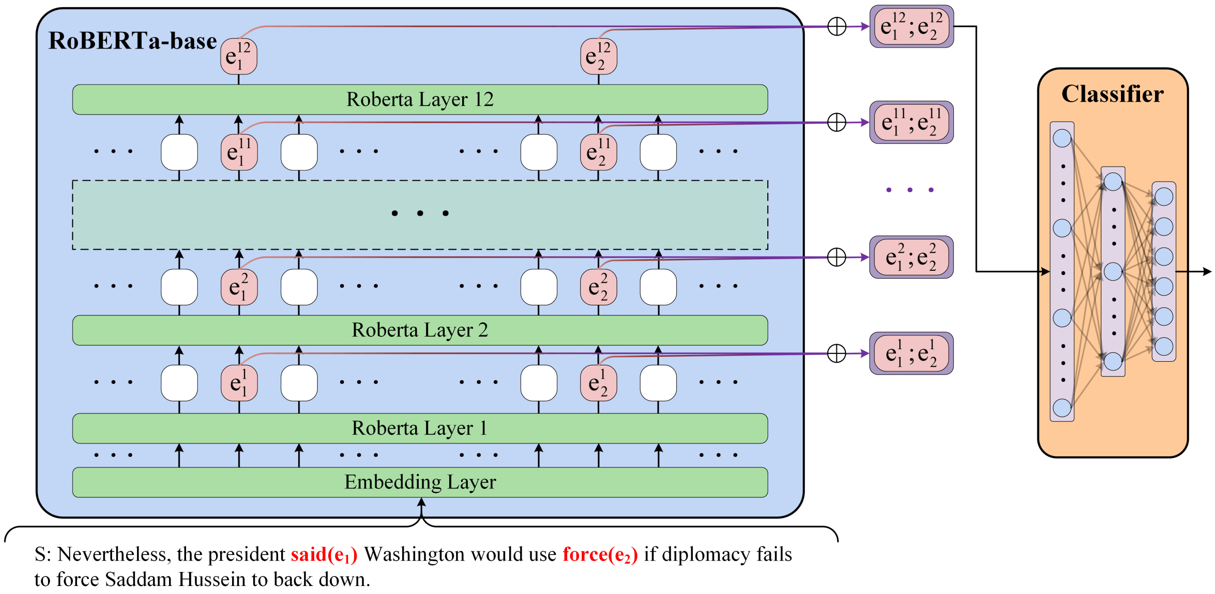

Figure 1: TempRel Classification model consisting of a pre-trained language model and a classifier.

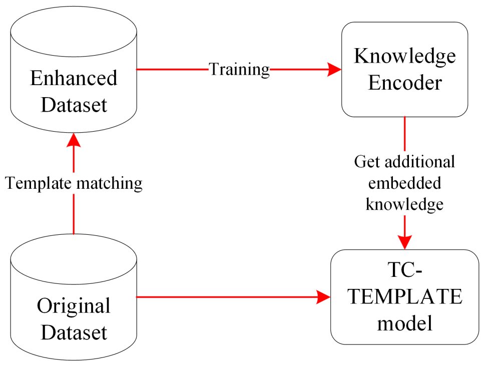

Figure 2: Process of training TempRel Classification model with embedded knowledge of TempRel information

label information. To solve this problem we design a fine-grained template in subsection 4.1. Another is how to learn embedded knowledge of TempRel information better. To solve this problem we design a more effective loss function in subsection 4.3. Figure 2 shows the overall process of our approach.

Firstly, we use templates to enrich the TempRel label information. We get an additional sentence set $S_{additional}$ through matching golden TempRel labels and event pairs in Data-set. Then connect each sentence s in original sentence set S and sentence s' in $S_{additional}$ to get a new sentence set S_{new}, which forms Enhanced Data-set together with original labels and event pairs. Secondly, we train the TempRel Classification model in section 3 with the Enhanced Data-set as Knowledge Encoder, then feed all sentences in S_{new} into the Knowledge Encoder to get all hidden states of event pairs of all intermediate layers in RoBERTa as additional embedded knowledge. Finally, we use both Original Data-set and additional embedded knowledge to train the TempRel Classification model. We name our TempRel Classification model Trained with EMbedded temPoral reLATion knowledgE as TC-TEMPLATE model. And we also name our approach as TEMPLATE.

Below are all the details of the components of our approach.

4.1 Build Templates

For taking full advantage of the TempRel label information, we aim to create effective templates which can automatically convert each golden TempRel label into a temporal information-enriched sentence s' to enrich golden TempRel label information. So, we design two kinds of templates with different granularity. For coarse-grained templates, we directly use the golden TempRel label and the event pair to describe the TempRel of event pair. For fine-grained templates, we claim that the time span of events (i.e. the duration of the events) guides TempRel classification. So we use the event start times, event end times and the TempRel between different events to describe the TempRel of the event pair in a more subtle level. We show both the coarse-grained templates and the fine-grained templates in Table 1.

TempRel	Coarse-Grained Templates	Fine-Grained Templates
AFTER*	the event of e_1 happens after event of e_2 happens.	the beginning of the event of e_1 is after the end of the event of e_2.
BEFORE*	the event of e_1 happens before event of e_2 happens.	the end of the event of e_1 is before the beginning of the event of e_2.
INCLUDES	the event of e_2 happens during the event of e_1 happens.	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 event of e_1 happens during the event of e_2 happens.	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.	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 happens simultaneously.	the event of e_1 and the event of e_2 have the same beginning and end time.

Table 1: Coarse-Grained Templates and Fine-Grained Templates. All the six TempRel labels are in TB-Dense and * indicates that the TempRel label also exists in MATRES.

4.2 Embedded Knowledge of TempRel Information

Having obtained suitable templates, we next convert the TempRel label information enriched by the templates into embedded knowledge which is more convenient for the TempRel Classification model to learn.

For each record (s, e_1, e_2, r) in data-set, we use r to match the templates and get s', then concatenate s and s' to get a new sentence $s_{new} = s; s'$, finally get a new record (s_{new}, e_1, e_2, r). We combine all new records into a new data-set i.e. Enhanced Dataset. We use the Enhanced Data-set to train TempRel Classification model as Knowledge Encoder, then use it to extract the embedded knowledge k = $\{\hat{e}_1^1;\hat{e}_2^1, \hat{e}_1^2;\hat{e}_2^2, \cdots, \hat{e}_1^{12};\hat{e}_2^{12}\}$ of TempRel information of each record. $\hat{e}_i^j$ is the hidden state corresponding to the event i from the j-th RoBERTa Layer, and $\hat{}$ is used to denote the hidden state come from Knowledge Encoder. Additionally, in the process of training Knowledge Encoder, we add a dropout layer between the RoBERTa and the Classifier, in order to make embedded knowledge k contain more useful temporal information.

4.3 Train the Model with Embedded Knowledge of TempRel Information

(Sun et al., 2019) prove the effectiveness of using mean-square error(MSE) loss between the normalized hidden states in distillation tasks as additional training loss. In this way, they make the small model learn from the large model. Motivated by this, we can make the event pair hidden states of all intermediate layers of TempRel Classification model $\{e_1^j; e_2^j\}_{j=1}^{12}$ to imitate the embedded knowledge$\{\hat{e}_1^j; \hat{e}_2^j\}_{j=1}^{12}$ by using L_{MSE} in equation 1. e_i^j is the hidden state corresponding to the event i from the j-th RoBERTa Layer. In this way, we enable TempRel Classification model to learn from embedded knowledge.

$$L_{MSE} = \sum_{j=1}^{12} \sum_{k=1}^{N} \left(\frac{1}{12 \times 2d} \cdot \left\| \frac{e_1^j; e_2^j}{\left\| e_1^j; e_2^j \right\|_2} - \frac{\hat{e}_1^j; \hat{e}_2^j}{\left\| \hat{e}_1^j; \hat{e}_2^j \right\|_2} \right\|_2^2 \right)_k \tag{1}$$

where N is the number of training samples, d is the hidden state dimension of RoBERTa.

Furthermore, we argue that the event hidden states of different intermediate layers are the event features under different perspectives and they have different importances to the learning process of the TempRel Classification model. But L_{mse} treat them as equally important. We also argue that the farther the embedded knowledge of $\hat{e}_1^j; \hat{e}_2^j$ is from the embedding of $e_1^j; e_2^j$, the more knowledge is contained in $\hat{e}_1^j; \hat{e}_2^j$ for a given layer j. So we propose a new method to automatically assign different weights $w_j \in \{ w_1, w_2, \cdots, w_{12} \}$ to each layer MSE loss. We design w_j as:

$$w_j = \frac{1 - \cos(\hat{e}_1^j; \hat{e}_2^j, e_1^j; e_2^j)}{\sum_{i=t}^{12}(1 - \cos(\hat{e}_1^t; \hat{e}_2^t, e_1^t; e_2^t))} \qquad (2)$$

We use the cosine values of the hidden states and the embedded knowledge of different intermediate layers to weight their importances. Then we get a new Cosine balanced MSE (C-MSE) loss L_{CMSE}:

$$L_{CMSE} = \sum_{j=1}^{12}\sum_{k=1}^{N}\left(\frac{w_j}{2d} \cdot \left\| \frac{e_1^j; e_2^j}{\left\|e_1^j; e_2^j\right\|_2} - \frac{\hat{e}_1^j; \hat{e}_2^j}{\left\|\hat{e}_1^j; \hat{e}_2^j\right\|_2} \right\|_2^2 \right)_k \qquad (3)$$

Combined with the cross entropy loss and the C-MSE loss, the final loss function can be formulated as:

$$L_{finall} = \alpha L_{CE} + \beta L_{CMSE} \qquad (4)$$

where α and β are hyper-parameters which weight the importances of discrete TempRel label information and additional enriched embedded knowledge of TempRel label information of intermediate layers.

5 Experiments and Results

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

5.1 Data-set

TB-Dense (Cassidy et al., 2014) is a densely annotated data-set for TempRel extraction, with 10 times as many relations per document as the TimeBank. It contains 6 types of relations: AFTER, BEFORE, INCLUDE, IS INCLUDED, VAGUE, SIMULTANEOUS. We use the same train (22 documents), dev (5 documents) and test (9 documents) splits as previous studies (Han et al., 2021; Zhang et al., 2021).

MATERS (Ning et al., 2018) contains 275 news documents from TimeBank (TB), AQUAINT (AQ), and Platinum (PT). It was annotated by a novel multi-axis annotation scheme with only 4 types of temporal relations: BEFORE, AFTER, EQUAL[3] and VAGUE. We follow the official split (i.e., TB+AQ (255 documents) as the train data-set and PT (20 documents) as the test data-set). As for the dev data-set, we use the same split strategy as previous studies (Ning et al., 2019; Tan et al., 2021). We randomly select 51 documents (20% of the official training data) as the dev data-set.

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

Corpora		Document	Tempral
TB-Dense	Train	22	4032
	Dev	5	629
	Test	9	1427
MATRES	Train	204	10097
	Dev	51	2643
	Test	20	837

Table 2: Data statistics for TB-Dense and MATRES

5.2 Experimental Setup

We use RoBERTa as our pre-trained language model for fine-tuning and optimize TempRel Classification model with BERTAdam. In the process of training Knowledge Encoder, we set the drop probability of the dropout layer between the RoBERTa and the Classifier to 0.5, in order to make the embedded knowledge contain more useful temporal information. We use grid search strategy to select best hyper-parameters, and select learning rate of Classifier $\in$ {1e-3, 5e-4}, learning rate of RoBERTa $\in$ {1e-5, 5e-6, 1e-6}, $\alpha \in$ [1.2: 0.7], $\beta \in$ [1200: 700] and batch size $\in$ {16, 24}. Since there are so many hyper-parameters to select for our approach, thus we first fix α = 1 and β = 1000, to search the best batch size and learning rates of Classifier and RoBERTa, then we fix them to search best α and β. As for the dimension of the hidden states between two fully connected layers in the Classifier, we set it to 36 and 16 for TB-Dense and MATRES respectively. The training time for one epoch takes about one minute on GeForce RTX 3090.

5.3 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

[3] We consider EQUAL to be the same as SIMULTANEOUS.

Model		TB-Dense	MATRES
JCL (Wang et al., 2020)	RoBERTa base	-	78.8
ECONET(Han et al., 2021)	RoBERTa Large	66.8	79.3
TGT (Zhang et al., 2021)	BERT Large	66.7	80.3
Poincaré Event Embeddings (Tan et al., 2021)	RoBERTa base	-	78.9
HGRU+knowledge (Tan et al., 2021)	RoBERTa base	-	80.5
TC-TEMPLATE$_{CMSE}$ (ours)	RoBERTa base	**69.07**	**81.97**
TC (ours)	RoBERTa base	63.56	79.95

Table 3: Comparison of various approaches on TempRel classification on TB-Dense and MATRES. Bold denotes the best performing model. TC denotes TempRel Classification model trained with only cross entropy loss. TC-TEMPLATE$_{CMse}$ denotes TempRel Classification model trained with additional embedded knowledge of fine-grained templates and C-MSE loss. F_1-score (%)

F_1 value for each model. The results of compared methods are directly taken from the cited papers. Next, we will introduce our comparison methods:

Joint Constrained Learning (JCL) conducts joint training on both temporal and hierarchical relation extraction based on RoBERTa and Bi-LSTMs. Meanwhile it uses logical constraints and common sense knowledge. **ECONET** uses a continual pre-trained method with mask prediction and contrastive loss to further train pre-trained language model with a large-scale TempRel corpus. **TGT** utilizes syntactic graph and designs a new Temporal Graph Transformer network. **Poincaré Event Embeddings** leverages hyperbolic embeddings to directly infer event relations through simple classifier. **HGRU** embeds events into hyperbolic spaces, devises an end-to-end architecture composed of hyperbolic neural units and introduces common sense knowledge.

We observe that our TC model achieves $63.56\% F_1$ on TB-Dense and $79.95\% F_1$ on MATRES. It demonstrates that our TC model can effectively classify TempRel, and even achieves a competitive performance which is close to current best $80.5\% F_1$ on MATRES, although MATRES is more simple data-set. Furthermore, our TC-TEMPLATE$_{CMSE}$ outperforms previous top state-of-the-art method on TempRel classification with up to $2.27\% F_1$ on TB-Dense and $1.47\% F_1$ on MATRES. These experimental results well prove the effectiveness of the idea of encoding enriched golden TempRel label information and learning the embedded knowledge through C-MSE loss. There are two possible reasons for the effectiveness. One is that we not only take advantage of the large amount of semantic information contained in golden TempRel label which is lost by the traditional discrete one-hot labels, but also use templates to further enrich this information. At the same time, we obtain many target embedded knowledge with rich temporal information. The other reason is that our C-MSE loss function can make the TempRel Classification model learn from embedded knowledge better. C-MSE loss function forces the intermediate layer hidden states of events in the TempRel Classification model to imitate target embedded knowledge of golden TempRel label information. In this process, for better imitation, the TempRel Classification model force itself to extract more useful information related to TempRel from sentences which don't contain any golden TempRel label.

Unlike ECONET and TGT, which use larger pre-trained language model, nor TGT and HGRU, which use networks with complex structure followed RoBERTa base or BERT Large, our approach enables a smaller and simpler model which only contains a RoBERTa base and two full connected layers to achieve the state-of-the-art performance.

5.4 Ablation Study

We observe that, compared with the TC model, the TC-TEMPLATE$_{CMSE}$ model performs better. This confirms the effectiveness of our embedded knowledge learning approach, which results in an improvements of $5.51\% F_1$ and $2.02\% F_1$ on TB-Dense and MATRES respectively. To go a step further, we study the effects of our proposed C-MSE loss function and fine-grained templates through following ablation experiments.

Embedding knowledge learning by C-MSE vs MSE. In order to determine whether our proposed Cosine balanced MSE loss has a positive effect, we conduct a comparative experiment. Under the

Model	TB-Dense	MATRES
TC-TEMPLATE$_{CMSE}$	69.07	81.97
TC-TEMPLATE$_{MSE}$	67.87	81.51
TC	63.56	79.95

Table 4: Comparison of TC-TEMPLATE models using C-MSE loss and using MSE loss under the premise of using fine-grained templates on TB-Dense and MATRES. F_1-score (%)

Model	TB-Dense	MATRES
TC-TEMPLATE$_{fine}$	69.07	81.97
TC-TEMPLATE$_{coarse}$	67.72	81.27
TC	63.56	79.95

Table 5: Comparison of TC-TEMPLATE models using fine-grained templates and coarse-grained templates under the premise of using C-MSE loss on TB-Dense and MATRES. F_1-score (%)

premise of using fine-grained templates, we record the experimental results of the TC-TEMPLATE using C-MSE loss and the TC-TEMPLATE using MSE loss on TB-Dense and MATRES respectively. The results are showed in Table 4. We can see that the TempRel Classification model using C-MSE achieves $1.2\% F_1$ and $0.46\% F_1$ performance improvement over the same model using simple MSE respectively, which demonstrates the benefit of using the cosine value of the intermediate layer hidden states $e_1^j; e_2^j$ corresponding to the events and the target embedded knowledge $\hat{e}_1^j; \hat{e}_2^j$ to balance the MSE losses of different intermediate layers in TempRel Classification model. Because the traditional MSE loss function treats each intermediate layer as equally important. It is easy to emphasize secondary knowledge while ignoring primary knowledge in training process. While our C-MSE loss function can re-emphasize the primary knowledge to solve this problem.

Fine-grained templates vs Coarse-grained templates. In order to study the impact of different granularity templates on model performance, we also compare F_1 values of TC-TEMPLATE models with different granularity templates on TB-Dense and MATRES under the premise of using C-MSE loss function. We report the results in Table 5. We can see that the TC-TEMPLATE using fine-grained templates performs better than the TC-TEMPLATE using coarse-grained templates, and has $1.35\% F_1$ and $0.70\% F_1$ improvements on TB-Dense and MATRES respectively. The reason for these improvements is not just that fine-grained templates are longer and describe TempRel more precisely compared to coarse-grained templates. It's also because fine-grained templates use the start and end times of the events to describe the TempRel. This way makes fine-grained templates closer to the annotation rules and actual basis for judgment than coarse-grained templates. Through the proven embedded knowledge learning process, the knowledge of fine-grained templates can better enable the model to heuristically learn the TempRel of events and it orient the TempRel Classification model to extract more distinguishable features.

6 Conclusion

In recent years, the mainstream TempRel classification neural networks focus on using discrete values to represent temporal relation categories and using a single cross entropy loss function to train the model, but it can't fully utilize the potential of the model. So we propose a new approach which train the TempRel Classification model with C-MSE loss and embedded knowledge from fine-grained templates and golden TempRel labels. Extensive experimental results on TB-Dense and MATRES data-sets show that our approach makes TempRel Classification model gain a huge improvement and TC-TEMPLATE$_{CMSE}$ performs better then all previous state-of-the-art methods on TempRel classification tasks.

7 Future Work

There may be two possible limitations in our approach. First, the enriched embedded knowledge of golden TempRel label information may contain some noises which confuse the TempRel Classification model in the learning process. Second, we lack experiments to demonstrate the effectiveness of our approach on more complex models, even though we consider our proposed method can be adapted to more complex models. Because we don't yet know how to balance the importances of different components of a complex model without introducing more hyper-parameters. So in the future, we will further investigate how to reduce the noise of embedded knowledge, and further refine our approach so that the performances of complex models can also be improved.

References

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. 2021. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12929–12936.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. Forecastqa: A question answering challenge for event forecasting with temporal text data.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. EventPlus: A temporal event understanding pipeline. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 56–65, Online. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. 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)*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. 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)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. 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)*, pages 4323–4332, Hong Kong, China. Association for Computational Linguistics.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385.

Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. Multi-task label embedding for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4545–4553, Brussels, Belgium. Association for Computational Linguistics.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. 2021. Extracting temporal event relation with syntactic-guided temporal graph transformer. *arXiv preprint arXiv:2104.09570*.

Xinyu Zhao, Shih-Ting Lin, and Greg Durrett. 2021. Effective distant supervision for temporal relation extraction. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 195–203, Kyiv, Ukraine. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Yichao Zhou, Yu Yan, Rujun Han, J Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14647–14655.