Dynamically Updating Event Representations for Temporal Relation Classification with Multi-category Learning

Fei Cheng[1], Masayuki Asahara[2], Ichiro Kobayashi[3], and Sadao Kurohashi[1]

[1]Graduate School of Informatics, Kyoto University
[2]National Institute for Japanese Language and Linguistics
[3]Ochanomizu University
{*feicheng,kuro*}*@i.kyoto-u.ac.jp, masayu-a@ninjal.ac.jp, koba@is.ocha.ac.jp*

Abstract

Temporal relation classification is a pair-wise task for identifying the relation of a temporal link (TLINK) between two mentions, i.e. event, time and document creation time (DCT). It leads to two crucial limits: 1) Two TLINKs involving a common mention do not share information. 2) Existing models with independent classifiers for each TLINK category (E2E, E2T and E2D) [1] hinder from using the whole data. This paper presents an event centric model that allows to manage dynamic event representations across multiple TLINKs. Our model deals with three TLINK categories with multi-task learning to leverage the full size of data. The experimental results show that our proposal outperforms state-of-the-art models and two transfer learning baselines on both the English and Japanese data.

1 Introduction

Reasoning over temporal relations relevant to an event mentioned in the document can help us understand when the event begins, how long it lasts, how frequent it is, and etc. Starting with the TimeBank (Pustejovsky et al., 2003) corpus, a series of temporal competitions (TempEval-1,2,3) (Verhagen et al., 2009, 2010; UzZaman et al., 2012) are attracting growing research efforts.

Temporal relation classification (TRC) is the task to predict a temporal relation (*after*, *before*, *includes*, etc.) of a TLINK from a source mention to a target mention. Less effort has been paid to explore the sharing information across 'local' pairs and TLINK categories. In recent years, a variety of dense annotation schemas are proposed to overcome the 'sparse' annotation in the original Timebank. A typical one is the Timebank-Dense (TD) corpus (Chambers et al., 2014), which performs a compulsory dense annotation with the complete graph of TLINKs for the mentions located in two neighbouring sentences. Such dense annotation increases the chance of pairs sharing common events and demands of managing 'global' event representations across pairs among TLINK categories.

However, globally managing event representations of a whole document takes an extremely heavy load for the dense corpora. Timebank-Dense contains around 10,000 TLINKs in only 36 documents and is 7 times denser than the original Timebank. Thus, we propose a simplified scenario called Source Event Centric TLINK (SECT) chain. For each event e_i in a document, we group all TLINKs containing the common source event e_i into the e_i centric TLINK chain and align them with the chronological order of the target mentions appearing in the document. We assume that our system is capable of learning dynamic representations of the centric event e_i along the SECT chain via a 'global' recurrent neural network (RNN).

> DCT: 1998-02-27
> *An intense* ***manhunt*** *(**e_1**) conducted by the FBI and the bureau of alcohol, tobacco and firearms* ***continues*** *(**e_2**) for Rudolph in the wilderness of western north Carolina. And* ***this week*** *(**t_1**), FBI director Louie Freeh assigned more agents to the* ***search*** *(**e_3**).*

We demonstrate our proposal with the above adjacent-sentence excerpt in Timebank-Dense. '(e_s, e_t)' denotes a directed TLINK from the source e_s to target e_t in this paper. Considering the '***manhunt*** (e_1)' centric chain: $\{(e_1, DCT), (e_1, e_2), (e_1, t_1), (e_1, e_3)\}$[2], '***manhunt***' holds a '*includes*' relation to '***continues***'.

[1]Time-to-Time (T2T) is not included in this paper, as we focus on event centric representations.

[2]As DCT is not explicitly mentioned in documents, we always place (e_i, DCT) on the top of a SECT chain

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1352–1357
November 16 - 20, 2020.

We assume that dynamically updating the representation of '***manhunt***' in the early step '(e_1, e_2)' will benefit the prediction for the later step (e_1, e_3) to '***search***'. '***manhunt***' is supposed to hold the same '*includes*' relation to '***search***', as the search should be included in the continuing manhunt.

Our model further exploits a multi-task learning framework to leverage all three categories of TLINKs in the SECT chain scope. A common BERT (Devlin et al., 2019) encoder layer is applied to retrieve token embeddings. The global RNN layer manages the dynamic event and TLINK presentations in the chain. Finally, our system feeds the TLINK representations into their corresponding category-specific (E2D, E2T and E2E) classifiers to calculate a combined loss.

The contribution of this work is listed as follows: 1) We present a novel source event centric model to dynamically manage event representations across TLINKs. 2) Our model exploits a multi-task learning framework with two common layers trained by a combined category-specific loss to overcome the data isolation among TLINK categories. The experimental results suggest the effectiveness of our proposal on two datasets. All the codes of our model and two baselines is released. [3]

2 Related Work

2.1 Temporal Relation Classification

Most existing temporal relation classification approaches focus on extracting various features from the textual sentence in the local pair-wise setting. Inspired by the success of neural networks in various NLP tasks, Cheng and Miyao (2017); Meng et al. (2017); Vashishtha et al. (2019); Han et al. (2019b,a) propose a series of neural networks to achieve accuracy with less feature engineering. However, these neural models still drop in the pair-wise setting.

Meng and Rumshisky (2018) propose a global context layer (GCL) to store/read the solved TLINK history upon a pre-trained pair-wise classifier. However, they find slow converge when training the GCL and pair-wise classifier simultaneously. Minor improvement is observed compared to their pair-wise classifier. Our model is distinguished from their work in three focuses: 1) We constrains the model in a reasonable scope, i.e. SECT chain. 2) We manages dynamic event representations, while their model stores/reads pair history 3) Our model integrates category-specific classifiers by multi-task learning, while they use the categories as the features in one single classifier.

[3] https://github.com/racerandom/NeuralTime

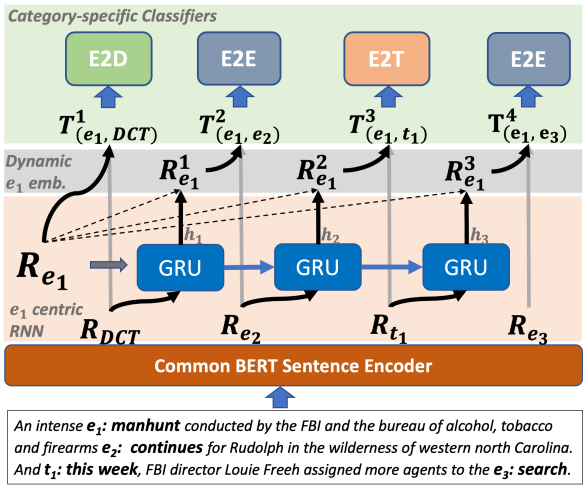

Figure 1: The overview of the proposed model.

2.2 Multi-task Transfer Learning

For the past three years, several successful transfer learning models (ELMO, GPT and BERT) (Peters et al., 2018; Radford et al.; Devlin et al., 2019) have been proposed, which significantly improved the state-of-the-art on a wide range of NLP tasks. (Liu et al., 2019) propose a single-task batch multi-task learning approach over a common BERT to leverage a large mount of cross-task data in the fine-tuning stage.

In this work, our model deals with various categories of TLINKs (E2E, E2T and E2D) in a batch of SECT chains to calculate the combined loss with the category-specific classifiers.

2.3 Non-English Temporal Corpora

Less attention has been paid for non-English temporal corpora. Until 2014, Asahara et al. starts the first corpus-based study BCCWJ-Timebank (BT) on Japanese temporal information annotation. We explore the feasibility of our model on this Japanese dataset.

3 Overview of Proposed Model

Figure 1 demonstrates the overview of our Source Event Centric (SEC) model with the previous e_1 centric chain example $\{(e_1, DCT), (e_1, e_2), (e_1, t_1), (e_1, e_3)\}$ in § 1.

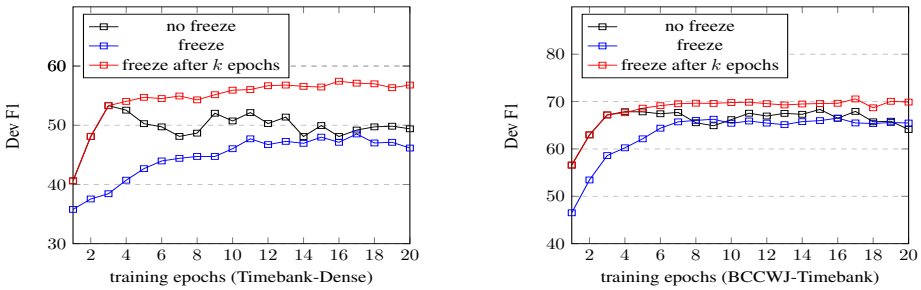

Figure 2: Dev performance (micro-F1) of three training strategies on two datasets.

3.1 BERT Sentence Encoder

We apply a pre-trained BERT for retrieving token embeddings of input sentences. For a multiple-token mention, we treat the element-wise sum of token embeddings as the mention embedding.

3.2 Source Event Centric RNN

After the BERT layer processing, the system collects all the mention embeddings appearing in the chain: $\{R_{e_1}, R_{DCT}, R_{e_2}, R_{t_1}, R_{e_3}\}$[4].

Our model assigns a 'global' two-layer gated recurrent unit (GRU) model with the left-to-right direction to simulate the chronological order of the SETC chain for updating the centric e_1 embeddings. The original e_1 embedding R_{e_1} is sent into the GRU as the initial hidden. At i-th TLINK step, the system inputs the target mention embedding to update the i-th e_1 embedding $R^i_{e_1}$ for generating the $\{i+1\}$-th step TLINK embedding T^{i+1}. As shown in Figure 1, the 3-rd TLINK embedding $T^3_{(e_1,t_1)}$ is the concatenation of the 2-nd step $R^2_{e_1}$ and target embedding R_{t_1} as the follows:

$$R^2_{e_1} = max(R_{e_1}, GRU(R_{e_2}, h_1)) \quad (1)$$

$$T^3_{(e_1,t_1)} = [R^2_{e_1}; R_{t_1}] \quad (2)$$

The element-wise max is desiged to set the initial R_{e_1} as an anchor to avoid the quality dropping of new hiddens after long sequential updating.

3.3 Multi-category Learning

After obtaining all the TLINK embeddings $\{T^1_{(e_1,DCT)}, T^2_{(e_1,e_2)}, T^3_{(e_1,t_1)}, T^4_{(e_1,e_3)}\}$ in the SECT chain via the previous two common layers, the system feeds them into the corresponding category-specific classifiers. Each classifier is built with one linear full-connected layer and Softmax layer. The system calculates the combined loss as the follows to perform multi-category learning.

$$L = L_{E2E} + L_{E2T} + L_{E2D} \quad (3)$$

[4]As DCT is not explicitly mentioned in documents, we set R_{DCT} as a trainable embedding.

Corpus	E2D	E2T	E2E	MAT	SECT
English	1,494	2,001	6,088	-	5.5
Japanese	2,873	1,469	1,862	776	2.4

Table 1: Number of TLINKs in the English and Japanese corpora. 'SECT' denotes the average TLINK number per SECT chain. 'MAT' is defined in § 4.3

4 Experiments and Results

We conduct the experiments of applying the SEC model on both the English TD and Japanese BT corpora. Juman++ (Tolmachev et al., 2018)[5] is adopted to do morphological analysis for Japanese text. TD annotation adopts a 6-relation set (*after*, *before*, *simultaneous*, *includes*, *is_included* and *vague*). We follow the 'train/dev/test' data split[6] of the previous work. For BT, we follow a merged 6-relation set as (Yoshikawa et al., 2014). We perform the document-level 5-fold cross-validation. In each split, we randomly select 15% documents as the dev set from the training set. The TLINKs statistics of the two corpora are listed in Table 1.

We adopt the English and Japanese pre-trained 'base' BERT[7] and empirically set RNN hidden size equal to BERT hidden, 4 SECT chains per batch, 20 epochs, and AdamW (lr=5e-5). The other hyper-parameters are selected based on the dev micro-F1. All the results are 5-run average.

For the lack of comparable transfer learning approaches, we build two BERT baselines as follows (fine-tuning 5 epochs, batch size is 16):

- **Local-BERT**: The concatenation of two mentions as TLINK embeddings are fed into the independent category-specific classifier.
- **Multi-BERT**: The multi-category setting as (Liu et al., 2019) of Local-BERT. Each time the system pops out a single-category batch,

[5]https://github.com/ku-nlp/jumanpp
[6]www.usna.edu/Users/cs/nchamber/caevo
[7]github.com/huggingface/transformers

encodes it via the common BERT, and feed it to the category-specific classifier.

'Local-BERT' and 'Multi-BERT' serve as the baselines in the ablation test for the proposed 'SEC' model. 'Local-BERT' is the 'SEC' model removing both global RNN and multi-category learning. 'Multi-BERT' is viewed as the 'SEC' model removing global RNN.

4.1 Asynchronous Training Strategy

Fine-tuning BERT is difficultly performed with training SEC RNN simultaneously. The standard fine-tuning only requires 3 to 5 epochs, which indicates the pre-trained model tends to quickly overfit. However, the SEC RNN is randomly initialized and requires more training epochs.

- **no freeze** of BERT sentence encoder
- **freeze** of BERT sentence encoder
- **freeze after k epochs**

Figure 2 shows the validation micro F1 of all TLINKs against the training epochs of the above asynchronous training strategies. **no freeze** shows the evidence of our concern that the curve undulate after the initial 3 epochs. **freeze** performs a stable learning phase with the lowest initialization. **freeze after k epochs** achieves the balance of the stability and high F1. Therefore, we perform the third strategy for all the following experiments. The number k is selected from $\{3, 4, 5\}$ based on the validation scores.

4.2 Main Timebank-Dense Results

Table 2 shows the experimental results on the English TD corpus. 'CATENA' (Mirza and Tonelli, 2016) is the feature-based model combined with dense word embeddings. 'SDP-RNN' (Cheng and Miyao, 2017) is the dependency tree enhanced RNN model.'GCL' (Meng and Rumshisky, 2018) is the global context layer model introduced in § 2.1. 'Fine-grained TRC' Vashishtha et al. (2019) is the ELMO based fine-grained TRC model with only the E2E results reported.

It's not surprising that the proposed model substantially outperforms state-of-the-art systems, as the existing SOTA didn't exploit BERT yet. Therefore, we offer the ablation test with 'Local-BERT'(w/o multi-categories learning and global SEC RNN) and 'Multi-BERT' (w/o global SEC RNN) to investigate the benefits of our two contributions. The 'SEC' model obtains +3.2, +6.8, +5.2 F1 improvements compared to 'Local-BERT', which suggests the effectiveness of two main proposal. The 'SEC' model further outperforms 'Multi-BERT' by 3.6 gain of the majority category E2E, 1.0 gain of E2T and 0.7 gain of E2D, which indicates the impact of the global SEC RNN.

A main finding is that E2E obtains higher gains from 'global' contexts, compare to E2T and E2D. It matches the intuition that events are more globally contextualized and time expressions are usually more self-represented (e.g. normalized time values). E2D mainly requires contextual information from the single sentences by the BERT encoder. E2T takes less advantage of BERT, while multi-category training with E2E, E2D can significantly improves its performance.

Models	**E2D**	**E2T**	**E2E**
Majority Vote	32.3	40.6	47.7
local Models			
CATENA (2016)	53.4	46.8	51.9
SDP-RNN (2017)	54.6	47.1	52.9
Fine-grained TRC (2019)	-	-	56.6
Local-BERT	62.7	49.4	59.8
local + multi-category Models			
Multi-BERT	65.2	54.8	61.4
global + multi-category Models			
GCL (2018)	48.9	48.7	57.0
SEC (proposed)	**65.9**	**55.8**	**65.0**

Table 2: Temporal relation classification results (micro F1) on the English Timebank-Dense.

Models	**E2D**	**E2T**	**E2E**	**MAT**
Majority Vote	68.3	50.4	43.2	39.3
local Models				
Yoshikawa (2014)	75.6	55.7	59.9	50.0
Local-BERT	80.7	58.9	61.2	54.1
local + multi-category Models				
Multi-BERT	81.4	**61.0**	63.3	61.6
global + multi-category Models				
SEC (proposed)	**81.6**	60.7	**64.5**	**64.6**

Table 3: Temporal relation classification results (micro F1) on the Japanese BCCWJ-Timebank.

4.3 Results on Non-English Data

Table 3 shows the results in the Japanese corpus. Different from the TD annotation schema, BT specifies two E2E categories for fitting the Japanese language: 1) E2E: between two consecutive events, 2) MAT: between two consecutive matrix verb events.

The state-of-the-art system on BT is the feature-

based approach (Yoshikawa et al., 2014). The comparisons are similar to the English data. Our 'SEC' obtains the substantial improvements compared to their work and two BERT baselines. An interesting observation is that MAT TLINKs are usually inter-sentence located at the end of SECT chains, as Japanese is a 'SOV' language. The results indicate that long distance MAT suffers from the low-quality representations in the 'local' setting and benefits from 'global' representation more.

5 Conclusion

This paper presents a novel transfer learning based model to boost the performance of temporal information extraction task especially for densely annotated dataset. Our model can dynamically update event representations across multiple TLINKs in a Source Event Centric chain scope. Our model exploits a multi-category learning framework to leverage the total data of three TLINK categories. The empirical results show that our proposal outperforms the state-of-the-art systems and the ablation tests suggest the effectiveness of two main proposals. The Non-English experiments support the feasibility of our system on the Japanese data.

References

Masayuki Asahara, Sachi Kato, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. 2014. Bccwj-timebank: Temporal and event information annotation on japanese text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 3, September 2014*.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

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.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. 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, Qiang Ning, and Nanyun Peng. 2019b. 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)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yuanliang Meng and Anna Rumshisky. 2018. Context-aware neural model for temporal information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 527–536, Melbourne, Australia. Association for Computational Linguistics.

Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. 2017. Temporal information extraction for question answering using syntactic dependencies in an lstm-based architecture. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Copenhagen, Denmark. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2016. On the contribution of word embeddings to temporal relation classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2818–2828, Osaka, Japan. The COLING 2016 Organizing Committee.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Katsumasa Yoshikawa, Masayuki Asahara, and Ryu Iida. 2014. Estimating temporal order relation for bccwj-timebank. In *Proceedings of the Japanese Annual Conference on NLP*. (in Japanese).