Dependency Enhanced Contextual Representations for Japanese Temporal Relation Classification

Chenjing Geng[†]
[†]Ochanomizu University
geng.chenjing@is.ocha.ac.jp

Fei Cheng[‡]
[‡]Kyoto Univresity
feicheng@i.kyoto-u.ac.jp

Masayuki Asahara[§]
[§]NINJAL
masayu-a@ninjal.ac.jp

Lis Kanashiro Pereira[†]
[†]Ochanomizu University
kanashiro.pereira@ocha.ac.jp

Ichiro Kobayashi[†]
[†]Ochanomizu University
koba@is.ocha.ac.jp

Abstract

Recently, quite a few studies have been progressive for temporal relation extraction, which is an important work used in several natural language processing applications. However, less concentration has been paid to corpora of Asian languages. In this work, we explored the feasibility of applying neural networks to temporal relation identification in the non-English corpora, especially Japanese corpora, BCCWJ-TimeBank. We explored the strength of combining contextual word representations (CWR) such as BERT (Devlin et al., 2019) and shortest dependency paths (SDP) for Japanese temporal relation extraction. We carefully designed a set of experiments to gradually reveal the improvements contributed by CWR and SDP. The empirical results suggested the following conclusions: 1) SDP offers richer information for beating the experiments with only source and target mentions. 2) CWR significantly outperforms fastText. 3) In most cases, the model applied CWR + SDP + Fine-tuning achieves the best performance overall.

1 Introduction

Temporal relation extraction is the task to identify temporal relationships between pairs of mentions, namely temporal expressions, and events and is useful in various Natural Language Processing (NLP) applications, such as question answering, storytelling, text summarization, etc. Many studies have so far proposed methods to extract temporal relation from English corpora (Pustejovsky et al., 2003a; UzZaman et al., 2012; Cheng and Miyao, 2017; Dligach et al., 2017; Tourille et al., 2017; Lin et al., 2017; Lin et al., 2018; Lin et al., 2019; Lin et al., 2020), however, unlike English, there are few temporal relation identification studies in Asian languages, due to limited data resources. On the other hand, Asahara et al. (2013) proposed the first Japanese temporal information corpus, BCCWJ-TimeBank (see, section 3 for more detail). Yoshikawa et al. (2014) used this Japanese corpus for temporal relation identification, however, there is still much space to explore for the temporal relation identification in Japanese. We, therefore, attempt to tackle it with the Japanese corpora in this study.

Temporal relation extraction is a kind of study of relation extraction. Bunescu and Mooney (2005) showed that there is a shortest path between the two entities in the dependency structure of a sentence, called shortest dependency path (SDP), which works well to extract the relation between the two entities. Xu et al. (2015) applied SDP in the BiLSTM framework for relation extraction and got the state-of-the-art results of those days. The SDP was also used in the deep learning framework for temporal relation extraction (Cheng and Miyao, 2017; Cheng and Miyao, 2018; Jiang et al., 2019; Li et al., 2019) and provided good results.

Recently, contextual word embedding models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have been effectively used as contextual word representations. Differently from word embeddings such as word2vec (Mikolov et al., 2013)

or GloVe (Pennington et al., 2014), these methods compute the embeddings for a sentence on the fly by taking the context of a target mention into account (Reimers et al., 2019).

From such a background, in this study, we decide to explore the strength of combining the shortest dependency paths and contextual word representations (CWR) for temporal relation classification in Japanese. We use a BERT model pre-trained on the NINJAL Web Japanese Corpus (NWJC)[1], to retrieve the contextual word representations of tokens.

The contributions of our work are as follows:

1. We presented a deep neural network-based model for Japanese temporal relation identification.

2. We conducted experiments achieving detailed empirical comparisons of various combinations of using contextual/non-contextual embedding and SDP/non-SDP model and investigated which features work best to extract temporal relations.

3. The experimental results show that both contextualized word representations and syntactic dependency information contribute to temporal relation identification.

2 Related studies

2.1 Multilingual Temporal Information Corpora

Starting with TimeBank (Pustejovsky et al., 2003b) and other temporal information corpora (XUE, 2005), a series of competitions on temporal information extraction (TempEval-1,2,3) (Verhagen et al., 2009; Verhagen et al., 2010; UzZaman et al., 2012) have been growing research efforts. While the Spanish data is released as TempEval-3, less attention has been paid to the Asian languages, such as Japanese, Chinese, etc. Asahara et al. (2013) started the first corpus-based study on annotating Japanese temporal information in the "The Balanced Corpus of Contemporary Written Japanese (BCCWJ)" corpus, and released it as BCCWJ-Timebank (Asahara et al., 2013). BCCWJ corpus is a balanced text resource containing extensive samples of modern Japanese texts and covering wide genres such as general books, magazines, newspapers, business reports, etc. To our knowledge, there is not any deep neural network-based model for tackling temporal relation identification with any Japanese corpora yet.

2.2 Neural Temporal Relation Classification

In recent years, quite a few studies on temporal relation extraction using deep neural networks have been proposed. Dligach et al. (2017) and Lin et al. (2017) used convolutional neural networks for temporal relation extraction and proposed a method for representing time expressions with single pseudo-tokens for CNNs. They established a new state-of-the-art result for a clinical temporal extraction task. Tourille et al. (2017) used BiLSTM (Hochreiter and Schmidhuber, 1997) to identify the relation between medical events and/or temporal expressions with the THYME corpus, clinical notes in English from the Mayo Clinic. Lin et al. (2018) proposed a recurrent neural network (RNN) with multiple semantically heterogeneous embeddings within a self-training framework. To extract temporal relation from medical corpora, they used both word embeddings made from clinical data sources and general domain sources and showed that their proposed method could generalize to new clinical domain data and obtained state-of-the-art performance in an unsupervised domain adaptation setting. Galvan et al. (2018) adopt the tree-based LSTM-RNN model proposed by Miwa and Bansal (2016) to temporal relation extraction from clinical texts.

In recent years, a large number of existing studies use BERT or other contextual word embedding because BERT provides a breakthrough in natural language processing, significantly outperforming previous state-of-the-art models on temporal relation extraction tasks. Han et al. (2019) established baselines for event temporal relation extraction on two story narrative datasets related to event description and causal and temporal relation, and applied their BERT-based method to extract those relations and showed that BERT worked well to extract the relations. Lin et al. (2019) applied BERT to extract temporal relation aiming to build a sentence-agnostic framework based on the fact that BERT is trained on large quantities of arbitrary spans of contiguous

[1] https://www.ninjal.ac.jp/english/database/type/corpora/

text instead of sentences. They also applied the idea of one-pass encodings for multiple relations extraction (Wang et al., 2019) for temporal relation extraction and increased its efficiency and scalability (Lin et al., 2020).

On the other hand, temporal relation extraction has been studied as a kind of relation extraction study, therefore, there is another approach for temporal relation extraction originating from the conventional relation extraction method. Bunescu and Mooney (2005) showed that the shortest path between the two entities in the dependency structure of a sentence, called 'shortest dependency path (SDP; see section 4.1 for detail)', works well to extract the relation between those entities.

As the studies to use SDP for relation identification of entities, Xu et al. (2015) presented SDP-LSTM, a neural network to classify the relation of two entities in a sentence using heterogeneous information along with the SDP as input information, and achieved a high F1-score with SemEval 2010 relation classification task. Jiang et al. (2019) proposed a BiLSTM-CNN-Attention model based on semantic dependency graph to extract sentence features extracting the SDP between the two entities from the semantic dependency graph as the input to the model. The shortest path combines the structural and semantic features of the sentence, which contributes to distinguishing between positive and negative examples in multi-instance learning. Their experimental results obtained high precision, showing that the model is adept in extracting structural features and semantic features.

Cheng and Miyao (2017) borrowed a state-of-the-art method of those days of relation extraction (Xu et al., 2015) and firstly introduced a SDP-based LSTM model to the temporal relation classification task. Their model achieved comparable performance without using external resources. Li et al. (2019) also showed that SDP worked well for relation extraction , demonstrating the effectiveness of syntactic structures in deep learning-based relation extraction by showing their proposed method significantly improved the performances on clinical notes.

In this study, we propose a neural network model combining SDP and CWR for temporal relation extraction for the Japanese corpus, i.e., BCCWJ-TimeBank.

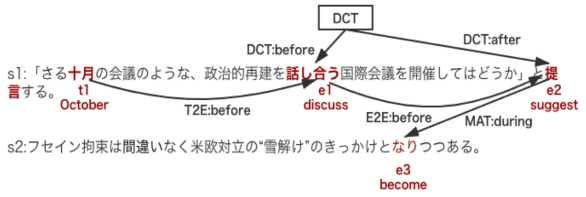


Figure 1: Example annotations in BCCWJ-TimeBank.

3 BCCWJ-TimeBank

The basic specifications of BCCWJ-TimeBank (Asahara et al., 2013) is based on TimeML (Pustejovsky et al., 2003a) and its temporal definition tags are adopted to Japanese language. There are four tags in the specification: <TIMEX3> for temporal expressions, <EVENT> and <MAKEINSTANCE> for event expressions, and <TLINK> for temporal ordering. As for <TLINK>, BCCWJ-TimeBank uses a variant of Allen's interval algebra (Allen, 1983); there are 13 labels for temporal ordering and three for event-subevent relations. Furthermore, it has a label ' vague ' for under-specified relations. So, we deal with 17 labels for temporal ordering. In our study, we focus on extracting temporal relations in the following four event relation types.

- **DCT**: Relation between an event instance and document creation time (DCT).
- **T2E**: Relation between a <TIMEX3>(non DCT) and an event instance within one sentence.
- **E2E**: Relation between two consecutive event instances.
- **MAT**: Relation between two consecutive matrix verbs of event instances.

Figure 1 shows an example of temporal relation extraction task in Japanese. There are three events (i.e. e1:話し合う (discuss), e2:提言 (suggest), e3:なり (become)), one time expression (i.e. t1:十月 (October)), and one DCT (Documentary Creation Time) in the sentence. The directed edges in the figure indicate the temporal relations between these entities. Table 1 shows the temporal relations in Figure 1.

Tasks	Temporal Relations		
DCT	DCT	BEFORE	e1
DCT	DCT	AFTER	e2
T2E	t1	BEFORE	e1
E2E	e1	BEFORE	e2
MAT	e2	DURING	e3

Table 1: Temporal relations <TLINK> in Figure 1

Original Labels	5+1 Labels	3+1 Labels
after	AFTER	AFTER
met-by	AFTER	AFTER
overlapped-by	AFTER-OVERLAP	OVERLAP
finishes	AFTER-OVERLAP	OVERLAP
during	OVERLAP	OVERLAP
started-by	OVERLAP	OVERLAP
equal	OVERLAP	OVERLAP
starts	BEFORE-OVERLAP	OVERLAP
contains	OVERLAP	OVERLAP
finished-by	OVERLAP	OVERLAP
overlaps	BEFORE-OVERLAP	OVERLAP
meets	BEFORE	BEFORE
before	BEFORE	BEFORE
is_included	OVERLAP	OVERLAP
identity	OVERLAP	OVERLAP
includes	OVERLAP	OVERLAP
vague	VAGUE	VAGUE

Table 2: Merging temporal relations into 5+1 and 3+1 labels

In our study, we prepared two temporal label sets, merging all of the 17 labels into 3+1 (AFTER, BEFORE, OVERLAP, and VAGUE) and 5+1 (AFTER, BEFORE, AFTER-OVERLAP, BEFORE-OVERLAP, and VAGUE) labels as shown in Table 2 in order to avoid the label sparsity problem as Cassidy et al. (2014) did.

As for statistical characteristics of BCCWJ-TimeBank corpora, Table 3 and 4 show human performance on temporal relation identification and the ratio of relation labels in the case of six labels, respectively. 'Agreement proportion' in Table 3 indicates the proportion of the unanimous annotations by three annotators in terms of annotating temporal relations on all data in BCCWJ-TimeBank. We see from the table that even manual annotation does not achieve a satisfying performance of temporal relation identification in Japanese, especially for Event-event (E2E) and cross-sentence link task (MAT), which proves that temporal relation identification is a challenging work. Table 4 shows the distribution of temporal relation labels in BCCWJ-TimeBank corpora in the case of six labels.

Tasks	# TLINKs	Agreement proportion
DCT	2854	74.3%
E2E	1642	55.2%
T2E	1513	69.1%
MAT	679	54.5%

Table 3: Agreement proportion of each TLINK task

Relation Label	DCT	T2E	E2E	MAT
AFTER	68.71 %	20.95%	26.45%	29.31%
BEFORE	20.04%	19.17%	44.79%	43.30%
OVERLAP	9.81%	49.70%	20.90%	20.32%
AFTER-OVERLAP	0.03%	2.24%	0	0
BEFORE-OVERLAP	0.07%	1.32%	0.06%	0
VAGUE	1.33%	6.60%	7.80%	7.07%

Table 4: Distribution of temporal relation labels in BCCWJ-TimeBank

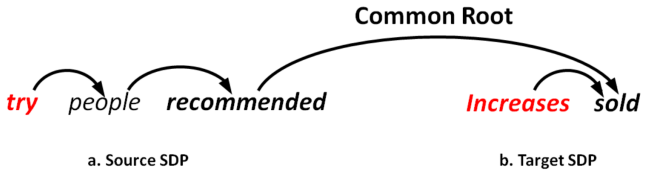
4 Temporal Relation Classifier

4.1 Shortest Dependency Path

Bunescu and Mooney (2005) applied the shortest dependency path (SDP) to relation extraction, based on the observation that the information required to assert a relationship between two named entities in the same sentence can be typically captured by the shortest path between the two entities in the dependency graph. The relation extraction method using the shortest dependency path (SDP), which contains the highly-covered words in the sentence, outperforms other methods such as using the whole sentence as input as pointed out by (Balali et al., 2020) because SDP can reduce the redundancy noise caused by needless information within a sentence.

Figure 3 shows an example of the shortest dependency in sentences. Each of the arrow points from modifier to its head. For example, word ' 改まっ (change)' and ' 大切 (cherish)' in S1 are connected by the words, ' 気分 (atmosphere)', ' し (do)'. So we see from the dependency graph that the dependency path between the two words is ' 改まっ(change), 気分 (atmosphere), し (do), 大切 (cherish)'.

We also follow the assumption proposed by Cheng and Miyao (2017) that there is a common root



S1: *For people who wants to **try** Noda's style, the Osechi set is recommended.*
S2: *"mini osechi" has been sold since three years ago, but the popularity **increases** gradually.*

Figure 2: Examples of a common root between two neighbouring sentences.

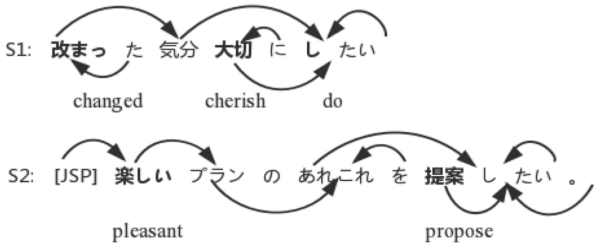


Figure 3: Examples of Short Dependency Path (SDP) in BCCWJ-TimeBank

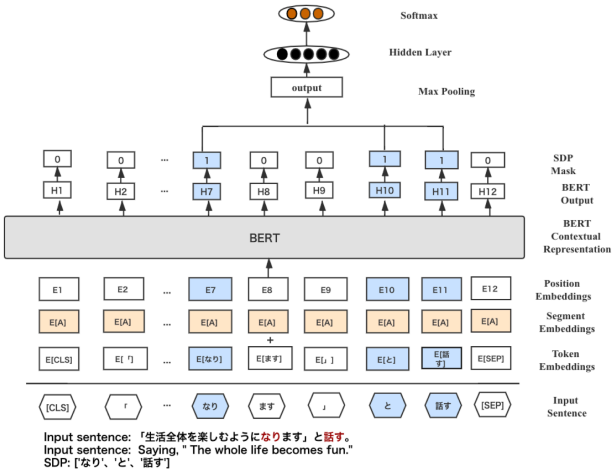


Figure 4: Temporal relation classifier

between the roots of two neighboring sentences so that a cross-sentence dependency path can be represented as the two shortest dependency path branches from the ends to the "common root" as shown in Figure 2 .

4.2 Model

We propose a neural network model for temporal relation extraction from Japanese corpora (see, Figure 4), which adopts BERT tokens pre-trained on the NWJC corpus as embedding input. An input sentence is represented with BERT tokens. Given a sentence, our model generates the shortest dependency path between source and target mentions, and then the tokens corresponding to SDP are selected as input to the classifier.

As for the DCT task, because there is only one entity in the DCT task, as the first step for the task, SDP between the event and the root of this sentence is obtained. Unlike DCT and the other two tasks, i.e., T2E and E2E, MAT is the task to extract temporal relation between two neighboring sentences, so we generate two shortest dependency paths of two sentences respectively and set the common root for the paths to extract temporal relation from the two sentences, following the idea by Cheng and Miyao (2017).

We have set two kinds of models in terms of using BERT: one is the model without fine-tuning, we call this model 'Feature-based model' and the other is the model with fine-tuning, we call this model 'Fine-tuning model'. In the case of using the Fine-tuning model, we do not freeze all the 12 layers of the BERT model, during the fine-tuning process.

5 Experiment and Results

5.1 Experiment Settings

We conducted several experiments for Japanese temporal relation classification on BCCWJ-TimeBank corpus (Asahara et al., 2013). As the data used in the experiments, we only use the temporal annotations agreed by all three annotators. In this study, we use document-level data split in order to avoid an overlap problem that words repeat in both training and

4 Labels Setting							
Tasks	Yoshikawa (2014)	fastText	fastText +SDP	BERT	BERT +Fine-tuning	BERT+SDP	BERT+SDP +Fine-tuning
DCT	N/A	65.3%	71.7%	75.2%	**83.4%**	74.7%	82.0%
E2E	N/A	56.6%	59.5%	53.1%	62.5%	54.5%	**62.7%**
T2E	N/A	46.8%	55.0%	47.7%	54.9%	52.7%	**55.3%**
MAT	N/A	44.5%	46.5%	47.2%	51.4%	51.0%	**52.5%**
6 Labels Setting							
Tasks	Yoshikawa (2014)	fastText	fastText +SDP	BERT	BERT +Fine-tuning	BERT+SDP	BERT+SDP +Fine-tuning
DCT	76%	65.4%	69.5%	74.0%	81.7%	75.8%	**81.9%**
E2E	60%	56.7%	59.0%	53.0%	61.6%	53.7%	**62.6%**
T2E	**54%**	43.7%	51.2%	45.0%	53.3%	48.4%	53.3%
MAT	50%	46.4%	45.9%	50.3%	50.8%	50.4%	**52.4%**

Table 5: Comparison of Experiments between Feature-based (Yoshikawa, 2014), FastText(+SDP), and BERT(+SDP, Finetuning)

test data. We use 5-fold cross-validation in our experiments. We randomly split all the 54 files into 5 folders, and use 4 folders as training and validation data, of which 15% is validation data. And the rest folder is used as test data. We conducted train and test experiments with 20 epochs, batch size is set as 16. The learning rates of Feature-based models (fastText model and BERT model) are 0.001, while it is 2e-5 in Fine-tuning model.

As the baseline to compare with the BERT model, we use the pre-trained Japanese word embeddings, nwjc2vec (Asahara, 2018) for fastText experiments, built from the NWJC corpus, and POS embeddings whose initial values are randomly decided with a lookup table of 50 dimensions. To compare our proposed model with the conventional model, we employed the model using a SDP-based BiLSTM for temporal relation extraction (Cheng and Miyao, 2017). The concatenation of POS and word embedding is fed into BiLSTM. We empirically set the dimension of the hidden layer of the BiLSTM as 300.

5.2 Experimental Results

The result of the experiments on identifying temporal relations with four and six labels on four tasks is shown in Table 5. We set the fastText model and BERT model as the baseline to compare with SDP models (fastText+SDP, BERT+SDP), in which the tokens related to SDP are used as input data instead of all the tokens in a sentence. The fastText+SDP model adopts word embeddings and POS embeddings of the words related to SDP as input information. While for the BERT+SDP model, BERT tokens related to SDP are used as input information.

Table 5 shows the experimental results with various model settings. In the experiments, we used embedding vectors at the last layer of BERT model. We used the Japanese time-order relationship estimation model proposed by (Yoshikawa et al., 2014) as a baseline for the task. They used support vector machine (SVM) to classify temporal order relations as a class classification problem and proposed an easily reproducible estimation method. Compared with the baseline, we found that the accuracy of the proposed method is improved in most cases except for the T2E task. FastText and BERT only use the source and target mentions as input, while the fastText+SDP and BERT+SDP use the word included in SDP as input.

The results of fastText and fastText+SDP were compared and the accuracy was improved in the case of using SDP in all tasks except for MAT task of six labels. We supposed that it is because when we classify the relations between two neighbouring sentences, there are cases where only source and target mentions are enough for input data. The results of both BERT and BERT+SDP were compared and the accuracy was improved in the case of using SDP in all tasks. In addition, the results of BERT+Fine-tuning and BERT+SDP+Fine-tuning were compared and the accuracy was improved in the case of using

Model Settings	BERT		BERT + Fine-tuning		BERT +SDP		BERT+SDP +Fine-tuning	
	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels
Last layer (with LSTM layer)	77.5%	77.5%	81.4%	80.9%	79.1%	78.9%	81.0%	81.1%
Last layer (without LSTM layer)	75.2%	74.0%	**83.4%**	81.7%	74.7%	75.8%	82.0%	**81.9%**
Concat all layer (with LSTM layer)	78.3%	**77.9%**	81.2%	80.9%	**79.7%**	**80.2%**	81.1%	81.0%
Concat all layer (without LSTM layer)	**78.6%**	76.2%	83.3%	**82.3%**	79.0%	78.7%	81.6%	81.6%

Table 6: Comparison of Experiments between several experiments settings on DCT task.

Model Settings	BERT		BERT + Finetuning		BERT +SDP		BERT+SDP +Fine-tuning	
	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels
Last Layer (with lstm layer)	60.5%	60.3%	62.3%	61.6%	**62.5%**	60.4%	61.7%	61.1%
Last Layer (without lstm layer)	53.1%	53.0%	62.5%	61.6%	54.5%	53.7%	**62.7%**	62.6%
Concat all Layer (with lstm layer)	**62.7%**	**62.5%**	60.5%	62.0%	62.0%	**63.4%**	60.0%	62.0%
Concat all Layer (without lstm layer)	55.9%	54.3%	**62.9%**	**63.3%**	55.9%	54.6%	61.7%	**63.1%**

Table 7: Comparison of Experiments between several experiments settings on E2E task.

SDP in all tasks except for DCT task in the four labels experiment. We assume the reason for this is that there is only one entity related to temporal event in a sentence in DCT task, therefore, it was not necessary to use the information of SDP.

In most cases, the results of four labels are higher than those of six labels because the experiment on four label identification is usually simpler than the six label identification. However, according to Table 5, the results of MAT task in six label identification are higher in the case of applying fastText model, fastText+SDP and BERT models. We assume that this is because, according to Table 4, there is no data with AFTER-OVERLAP or BEFORE-OVERLAP labels, so the data are the same between four labels and six labels in MAT task. In addition, if the results of fastText and BERT, fastText+SDP and BERT+SDP models were compared, the models with BERT were found to provide high performance in all cases, and BERT+SDP+Fine-tuning showed the highest accuracy in both four and six tasks experiments in most cases.

5.3 Further Study

In order to further explore the functionality of BERT model for temporal relation identification, we conducted evaluation experiments on the models with the case of using the information from multiple layers of BERT encoder as input information, and that of with or without LSTM layer after BERT encoder.

Table 6, 7, 8, 9 show the experimental results of input information through various settings of four tasks. The 'Last layer' indicates the case where embeddings from the last layer of BERT encoder are used, and 'Concat all layer' indicates the case where all layers of BERT encoder are concatenated and used. 'With LSTM layer' indicates the case where the embeddings of all the tokens of the last layer or the whole layers of BERT encoder are input to a BiLSTM. The BiLSTM is put in the model before the softmax layer for classification. While 'without LSTM layer' indicates the same model as shown in Figure 4.

In each table of results, to investigate the influence between LSTM model and BERT model, we com-

Model Settings	BERT		BERT + Finetuning		BERT +SDP		BERT+SDP +Fine-tuning	
	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels
Last Layer (with lstm layer)	55.7%	49.3%	54.6%	50.8%	57.0%	55.8%	57.1%	51.8%
Last Layer (without lstm layer)	47.7%	45.0%	54.9%	51.7%	52.7%	48.4%	55.3%	**53.3%**
Concat all Layer (with lstm layer)	53.7%	**52.2%**	60.5%	62.0%	**58.1%**	**53.8%**	55.0%	53.1%
Concat all Layer (without lstm layer)	**55.9%**	53.0%	**62.9%**	**63.3%**	53.7%	45.4%	**57.7%**	52.9%

Table 8: Comparison of Experiments between several experiments settings on T2E task.

Model Settings	BERT		BERT + Finetuning		BERT +SDP		BERT+SDP +Fine-tuning	
	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels	4 labels	6 labels
Last Layer (with lstm layer)	50.0%	**53.4%**	47.6%	47.4%	53.5%	51.1%	47.4%	47.8%
Last Layer (without lstm layer)	47.2%	50.3%	51.4%	50.8%	51.0%	50.4%	52.5%	52.4%
Concat all Layer (with lstm layer)	52.6%	53.3%	48.9%	47.3%	**54.6%**	**54.7%**	55.0%	**53.1%**
Concat all Layer (without lstm layer)	**50.0%**	50.3%	**54.5%**	**51.6%**	52.0%	52.7%	**57.7%**	52.9%

Table 9: Comparison of Experiments between several experiments settings on MAT task.

pared the models between with and without LSTM layer using the embeddings of the last layer of BERT encoder, and also the models between with and without LSTM layer using the embeddings of all layers of BERT encoder. In most cases of Feature-based models (BERT and BERT+SDP), the models without LSTM layer outperform the models with LSTM layer overall. In the case of the Fine-tuning models (BERT+Fine-tuning, BERT+SDP+Fine-tuning), the models without LSTM layer got better performance overall. We assume that there happens conflict between LSTM model and BERT model. This is an issue for the future research to further explore. In addition, when we extract CWR from all 12 layers of BERT encoder, we get higher accuracy than only last layer. We suppose that CWR from all 12 layers of BERT encoder provide richer information than that from only the last layer.

6 Conclusion

In this work, we explored the strength of combining contextual word representations (CWR) and shortest dependency paths (SDP) for Japanese temporal relation classification. We carefully designed a set of experiments to gradually reveal the improvements contributed by CWR and SDP. The empirical results suggested following conclusions: 1) SDP offers richer information for beating the baseline with only source and target words. 2) CWR significantly outperforms fastText. 3) CWR+SDP+Fine-tuning achieves the best performance overall. In the future work, we plan to investigate a further validation of the assumptions that we did in experimental results section, and future studies could investigate the association between LSTM model and BERT model.

Acknowledgments

This work was supported by JSPS KAKENHI Grants Number 18H05521 and 21H00308.

References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843,

November.

Masayuki Asahara, Sachi Yasuda, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. 2013. BCCWJ-TimeBank: Temporal and event information annotation on Japanese text. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 206–214, Taipei, Taiwan, November. Department of English, National Chengchi University.

Masayuki Asahara. 2018. Nwjc2vec: Word embedding dataset from ‘ ninjal web japanese corpus. *International Journal of Theoretical and Applied Issues in Specialized Communication*, 24.

Ali Balali, Masoud Asadpour, Ricardo Campos, and Adam Jatowt. 2020. Joint event extraction along shortest dependency paths using graph convolutional networks.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 724–731, Stroudsburg, PA, USA. 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, Baltimore, Maryland, June. Association for Computational Linguistics.

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, July. Association for Computational Linguistics.

Fei Cheng and Yusuke Miyao. 2018. Inducing temporal relations from time anchor annotation. 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 1833–1843, New Orleans, Louisiana, June. 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, June. Association for Computational Linguistics.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain, April. Association for Computational Linguistics.

Diana Galvan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. 2018. Investigating the challenges of temporal relation extraction from clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

Rujun Han, Mengyue Liang, Bashar Alhafni, and Nanyun Peng. 2019. Contextualized word embeddings enhanced event temporal relation extraction for story understanding. *CoRR*, abs/1904.11942.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Ming Jiang, Jiecheng He, Jianping Wu, Chengjie Qi, and Min Zhang. 2019. Relation extraction based on semantic dependency graph. *Journal of Computational Methods in Sciences and Engineering*, pages 1–12, 08.

Zhiheng Li, Zhihao Yang, Chen Shen, Jun Xu, Yaoyun Zhang, and Hua Xu. 2019. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. *BMC Medical Informatics and Decision Making*, 19(1), January.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP 2017*, pages 322–327, Vancouver, Canada,, August. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 165–176, Brussels, Belgium, October. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020. A BERT-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 70–75, Online, July. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. Timeml: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5*.

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

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A bi-LSTM approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230, Vancouver, Canada, July. 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*.

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.

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting multiple-relations in one-pass with pre-trained transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1371–1377, Florence, Italy, July. Association for Computational Linguistics.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794, Lisbon, Portugal, September. Association for Computational Linguistics.

N. XUE. 2005. The penn chinese treebank : Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11.

Katsumasa Yoshikawa, Masayuki Asahara, and Iida Ryu. 2014. Estimation of temporal ordering relation with bccwj-timebank (in japanese). In *Proceedings of the Japanese Annual Conference: 2014 The Association for Natural Language Processing(NLP 2014)*.