Dependency Enhanced Contextual Representations for Japanese Temporal Relation Classification

Chenjing Geng[†], Lis Kanashiro Pereira[†], Fei Cheng[‡], Masayuki Asahara[§], Ichiro Kobayashi[†]
Ochanomizu University[†],Kyoto University[‡],National Institute for Japanese Language and Linguistics[§]
{geng.chenjing, koba}@is.ocha.ac.jp,kanashiro.pereira@ocha.ac.jp,
feicheng@i.kyoto-u.ac.jp, masayu-a@ninjal.ac.jp

1 Introduction

Temporal relation identification is the task of identifying temporal relationships between pairs of entities, namely temporal expressions and events, and is useful in various Natural Language Processing (NLP) applications, such as question answering, story telling, text summarization, etc. Temporal relation identification is a challenging task, especially for Asian languages such as Japanese and Chinese. Figure 1 shows an example of temporal relation extraction task. There are three events (i.e. e1, e2, e3), one time expression (i.e. t1), and one DCT inside. The directed edges in the figure indicate the temporal relations between these entities. Table 1 shows the temporal relations in Figure 1.

Unlike English, there are few temporal relation identification studies in Japanese, due to limited data resources. Recently, contextual word embedding models, such as BERT[1] and ELMo [2] have been effectively used as contextual word representations, alleviating the need for additional resources when training a model. Differently from word embeddings such as word2vec [3] or GloVe [4], these methods compute the embeddings for a sentence on the fly by taking the context of a target word into account [5].

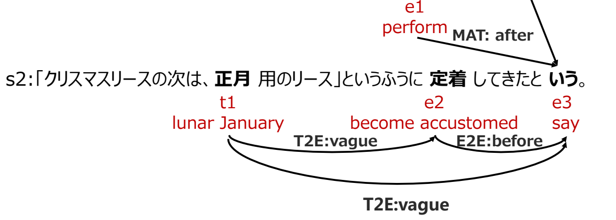
DCT
DCT : before
DCT : after
s1:室内に掛けられるリースは、手軽にお正月気分が 演出 できる。
e1
perform
MAT: after
s2:「クリスマスリースの次は、正月 用のリース」というふうに 定着 してきたと いう。
t1
lunar January
e2
become accustomed
e3
say
T2E:vague
E2E:before
T2E:vague

Figure 1: An example of <TLINK> in BCCWJ-TimeBank.

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

Table 1: Temporal relations in Figure 1

In this work, we explore the strength of combining contextual word representations (CWR) and shortest dependency paths (SDP) for Japanese temporal relation classification. We use a BERT model pre-trained on the NINJAL Web Japanese Corpus (NWJC), to extract contextual word representations. These representations are used as input to a bidirectional long short-term memory (BiLSTM). By comparing the results of our BERT model with a baseline using the non-SDP information word2vec model and SDP-based word2vec model, our results show that contextualized word embedding and dependency information contribute to temporal relation identification.

2 Related work

In recent years, quite a few studies on the temporal relation extraction have been proposed. Those include convolutional neural network (CNN) [6], whose work is experiment with convolutional neural network for temporal relation extraction and establish a new state-of-the-art for several scenarios, BiLSTM [7],which outperforms classical approaches based on feature engineering, etc. Among all works, LSTM shows its predomination on temporal relation extraction.

[8] proposed several neural network models processing SDP inputs for relation extraction in different text domains. [9] first introduced a SDP-based

LSTM model to the temporal relation classification task.

Their work achieved considerably good results without relying on feature extraction from external resources. In this work, we propose a neural model of combining SDP and contextual word representations for temporal relation classification in the Japanese BCCWJ-Timebank corpus.

3 BCCWJ-TimeBank

The basic specifications of BCCWJ-TimeBank[10] is based on TimeML [11] and its temporal definition tags are adopted to Japanese language.In our work, our target is divided into four tasks as following.

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

In our study, we have 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 in the experiments so that our labels can become close to those of TimeBank-Dense [12].

4 Temporal Relation Classifier

4.1 Shortest Dependency Path

[13] presented 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 typically captured by the shortest path between the two entities in the dependency graph. A shortest dependency path (SDP), which contains the highly-covered words in the sentence, outperforming other methods such as using the whole sentence as input which was mentioned in works such as [14] because it can reduce the redundancy noise caused by needless information within a sentence.

We follow the assumption proposed by [9] that there is a common root between the roots of two neighbouring 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”.

4.2 Model

We propose a neural Japanese temporal relation classification model, which adopts a BERT encoder (pretrained on the NWJC corpus) to embedding inputs. One of our main purposes is to investigate the qualities of different word representations provided by contextual models and traditional word2vec. To the end, we decide to freeze the BERT encoder during training. Otherwise, we can hardly distinguish the contribution of contextual representations or deep transformer layers of BERT.

Given a sentence, our model generates a shortest dependency path between a source and a target word. Because there is only one entity in DCT task, as the first step for the task, SDP between the event and the root of this sentence is obtained. Moreover, MAT shows the relations between two neighbouring sentences, so we assume that there is a common root between two neighbouring sentences as mentioned in [9] Then word embeddings are concatenated and fed to BiLSTM as input information.

We feed as input to BERT the lemmatized form of the original sentence and the SDP representation of the source and target entities. Following the BERT’s specific input format, a special token [SEP] is added between the tokens of the full sentence and the SDP. We extract the representation of all the tokens of second to the last layer of the pre-trained BERT model and feed them into a BiLSTM. The forward and backward outputs of the BiLSTM are concatenated, and fed into a fully connected hidden units layer. The softmax layer finally outputs a multi-class prediction for the temporal relations.

Figure 2 shows the temporal relation classifier for the task of DCT, E2E, T2E and MAT.

5 Experiment and Results

5.1 Experiment Settings

We conduct experiments for our Japanese temporal relation classification models on BCCWJ-TimeBank corpus [10]. For the annotation agreement, we only use the temporal annotations agreed by all three annotators. ’Agreement proportion’ in Table 2 indicates the proportion of the unanimous annotations by three annotators among all data. We see from the table that even manual annotation does not achieve a satisfied performance of temporal relation identifying in Japanese, especially for Event-event and cross-sentence link task (MAT), which proves that temporal relation identification is a challenging work. Since there are several temporal relation pairs between a particular word and the other words in the same document. In this study, we use document-level data

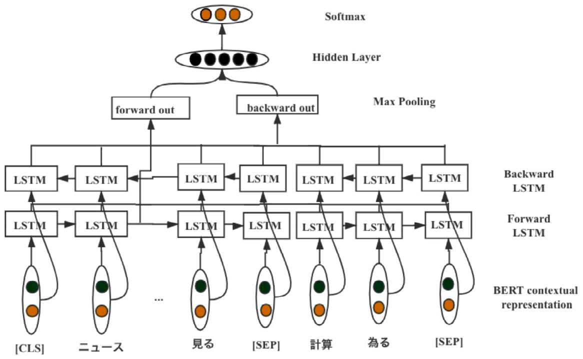

Figure 2: Temporal relation classifier

split in order to avoid an overlap problem that words repeat in both training and test data. We randomly select ten files from 54 files for test data, and the rest files are used as training data.We did our train and test experiments at 20 epochs.

As the baseline to compare with BERT model, we use the pre-trained Japanese word embeddings, nwjc2vec [15] for word2vec experiments, built from the NWJC corpus, and POS embeddings whose initial values are randomly decided with a lookup table of 50 dimensions. The concatenation of POS and word embedding is fed into BiLSTM. We empirically set the learning rate as 0.001 and the dimension of the hidden layer of the BiLSTM is set as 300.

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

Table 2: Agreement proportion of each TLINK task

5.2 Results

The result of the experiments on identifying temporal relations with four and six labels on four tasks is shown in Table 3. We set the word2vec model as the baseline, in which experiment the source and target words of the tasks are the only input data. The word2vec+SDP model adopt word embeddings and POS embeddings of SDP words of two entities. While for the BERT+SDP model, we extract embedding of SDP words from NWJC-BERT.

In Table 3, 'Majority vote' indicates the temporal relation with the most proportion among all four (or six) relations to be identified. By comparing the results of word2vec and word2vec+SDP, it is obviously that except for E2E task, the accuracy of the other three tasks rise apparently after SDP is adopted, which indicates the predomination of SDP method for temporal relation identification task. By comparing the results of word2vec+SDP and BERT+ SDP, BERT outperforming in most cases except for E2E task(3+1 labels).Moreover, BERT+SDP shows the best performance among three experiments.

6 Conclusion

In this work, we explore 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 word2vec. 3) CWR + SDP achieves the best performance overall. In the future work, we plan to investigate the fine-tuning of BERT model, as deep transformer layers have prove their power in a wide range of NLP tasks.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 2019. Association for Computational Linguistics.

Tasks	word2vec		word2vec+SDP		BERT+SDP		Majority Vote
	3+1 labels	5+1 labels	3+1 labels	5+1 labels	3+1 labels	5+1 labels	
DCT	55.6%	58.2%	69.5%	69.3%	**74.2%**	**72.5%**	after(68.7%)
T2E	50.4%	50.4%	54.7%	51.8%	**55.5%**	**54.4%**	before(44.7%)
E2E	**63.9%**	49.4%	60.6%	50.4%	56.9%	**57.1%**	overlap(49.7%)
MAT	34.4%	33.6%	46.1%	45.5%	**48.2%**	**47.4%**	before(43.3%)

Table 3: Comparison between accuracy of experiments on word2vec(non-sdp) model, word2vec+sdp model and BERT+sdp model.

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

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

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

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

[6] Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 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 2017. Association for Computational Linguistics.

[7] Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 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 2017. Association for Computational Linguistics.

[8] Zhiheng Li, Zhihao Yang, Chen Shen, Jun Xu, Yaoyun Zhang, and Hua Xu. 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), 2019.

[9] Fei Cheng and Yusuke Miyao. 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 2017. Association for Computational Linguistics.

[10] Masayuki Asahara, Sachi Yasuda, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. 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 2013. Department of English, National Chengchi University.

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

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

[13] Razvan C. Bunescu and Raymond J. Mooney. 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, 2005. Association for Computational Linguistics.

[14] Huiwei Zhou, Huijie Deng, and Jiao He. Chemical-disease relations extraction based on the shortest dependency path tree. In *Proceedings of the fifth BioCreative challenge evaluation workshop. Spain: BioCreative 2015*, pages 214–219, 2015.

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