# Classifying Temporal Relations by Bidirectional LSTM over Dependency Paths

**Fei Cheng** and **Yusuke Miyao**
Research Center for Financial Smart Data
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
{fei-cheng, yusuke}@nii.ac.jp

## Abstract

Temporal relation classification is becoming an active research field. Lots of methods have been proposed, while most of them focus on extracting features from external resources. Less attention has been paid to a significant advance in a closely related task: relation extraction. In this work, we borrow a state-of-the-art method in relation extraction by adopting bidirectional long short-term memory (Bi-LSTM) along dependency paths (DP). We make a "common root" assumption to extend DP representations of cross-sentence links. In the final comparison to two state-of-the-art systems on TimeBank-Dense, our model achieves comparable performance, without using external knowledge and manually annotated attributes of entities (class, tense, polarity, etc.).

## 1 Introduction

Recently, the need for extracting temporal information from text is motivated rapidly by many NLP tasks such as: question answering (QA), information extraction (IE), etc. Along with the TimeBank[1] (Pustejovsky et al., 2003) and other temporal information annotated corpora, a series of temporal evaluation challenges (TempEval-1,2,3) (Verhagen et al., 2009, 2010; UzZaman et al., 2012) are attracting growing research efforts.

Temporal relation classification is a task to identify the pairs of temporal entities (events or temporal expressions) that have a temporal link and classify the temporal relations between them. For instance, we show an event-event (E-E) link with 'DURING' type in *(i)*, an event-time (E-T) link with 'INCLUDES' type in *(ii)* and an event-DCT (document creation time, E-D) with 'BEFORE' type in *(iii)*.

*(i)* *There was no* ***hint*** *of trouble in the last* ***conversation*** *between controllers and TWA pilot Steven Snyder.*

*(ii)* *In Washington* ***today****, the Federal Aviation Administration* ***released*** *air traffic control tapes.*

*(iii)* *The U.S. Navy* ***has*** *27 ships in the maritime barricade of Iraq.*

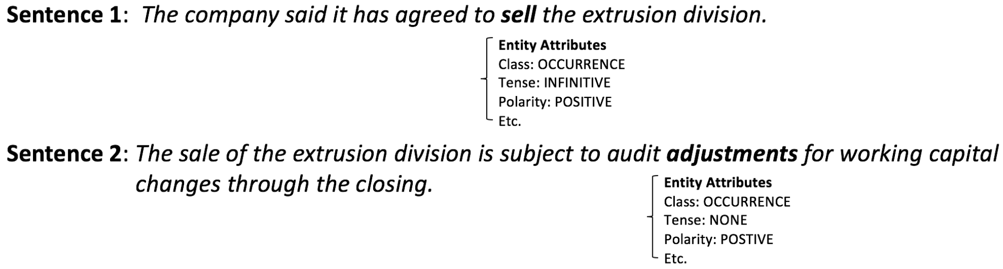
Marcu and Echihabi (2002) propose an approach considering word-based pairs as useful features. The following researchers (Laokulrat et al., 2013; Chambers et al., 2014; Mani et al., 2006; D'Souza and Ng, 2013) focus on extracting lexical, syntactic or semantic information from various external knowledge bases such as: WordNet (Miller, 1995) and VerbOcean (Chklovski and Pantel, 2004). However, these feature based methods rely on hand-crafted efforts and external resources. In addition, these works require the features of entity attributes (class, tense, polarity, etc.), which are manually annotated to achieve high performance. Consequently, they are hard to obtain in practical application scenarios.

In relation extraction, there is an explosion of the works done with the dependency path (DP) based methods, which employ various models along dependency paths (Bunescu and Mooney, 2005; Plank and Moschitti, 2013). In recent years, the DP-based neural networks (Socher et al., 2011; Xu et al., 2015a,b) show state-of-the-art performance, with less requirements on explicit features. Intuitively, the DP-based approaches have the potential to classify temporal relations.

Both relation extraction and temporal relation classification require the identification of relation-

[1] https://catalog.ldc.upenn.edu/LDC2006T08

*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 1–6
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2001


**Sentence 1**: *The company said it has agreed to* ***sell*** *the extrusion division.*

**Entity Attributes**
Class: OCCURRENCE
Tense: INFINITIVE
Polarity: POSITIVE
Etc.

**Sentence 2**: *The sale of the extrusion division is subject to audit* ***adjustments*** *for working capital changes through the closing.*

**Entity Attributes**
Class: OCCURRENCE
Tense: NONE
Polarity: POSTIVE
Etc.

Figure 1: An example of the sentences with entity attributes annotated in TimeBank.

ship between entities in texts. However, temporal relation classification is more challenging, since it includes three different type of entities: 'event', 'time expression' and DCT. Cross-sentence links also add additional complexity into the task. Due to the outstanding performance of DP-based neural networks revealed in relation extraction, we borrow this state-of-the-art approach to temporal relation classification.

In Section 2 of this paper, we review related work and introduce TimeBank-Dense. We discuss the cross-sentence link problem and the architectures of our E-E, E-T and E-D classifiers in Section 3. In Section 4, the experiments are performed on TimeBank-Dense and we compare our model to the baseline and two state-of-the-art systems. The final conclusion is made in Section 5.

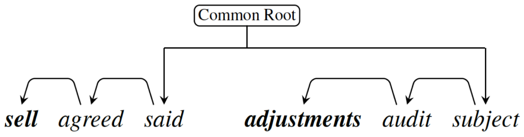
## 2 Background

### 2.1 Related Work

Current state-of-the-art temporal relation classifiers exploit a variety of features. Laokulrat et al. (2013); Chambers et al. (2014) extract lexical and morphological features derived from WordNet synsets. Mani et al. (2006); D'Souza and Ng (2013) incorporate semantic relations between verbs from VerbOcean as features. In addition, most of the systems include the entity attributes (Figure 1) specified in TimeML [2] as basic features, which actually need heavy human annotations.

In this work, we push this work into a more practical level by using only word, part-of-speech (POS), dependency parsing information, without incorporating entity attributes, as well as any other external resources.

In relation extraction, Bunescu and Mooney (2005) propose an observation that a relation can be captured by the shortest dependency path (SDP) between the two entities in the entire dependency graph. Plank and Moschitti (2013) extract syntactic and semantic information in a tree kernel. Following this line, researchers (Socher et al., 2011; Xu et al., 2015a,b) achieve state-of-the-art performance by building various neural networks over dependency path.



(a) The source SDP.
(b) The target SDP.

Figure 2: An example of the DP representation of a cross-sentence link between the two sentences in Figure 1.

Our system is similar to the work by Xu et al. (2015b). They perform LSTM with max pooling separately on each feature channel along dependency path. In contrast, our system adopts bidirectional LSTM on the concatenation of feature embeddings.

### 2.2 TimeBank-Dense

In the original TimeBank, temporal links have been created on those pairs with semantic connections, which led to a sparse annotation style. Cassidy et al. (2014) [3] propose a mechanism to force annotators to create complete graphs over the entities in neighboring sentences. Compared to 6,418 links in 183 TimeBank documents, TimeBank-Dense achieves greater density with 12,715 links in 36 documents.

We follow a similar experiment setting to the other two systems (Mirza and Tonelli, 2016; Chambers et al., 2014) with the same 9 documents

[2] http://timeml.org/

[3] https://www.usna.edu/Users/cs/nchamber/caevo

as test data and the others as training data (15% of training data is split as validation data for early stopping).

## 3 The Proposed Method

### 3.1 Cross-sentence Dependency Paths

Intuitively, the dependency path based idea can be introduced into the temporal relation classification task. However, around 64% E-E, E-T links in TimeBank-Dense are with the ends in two neighboring sentences, called cross-sentence links.

A crucial obstacle is how to represent the dependency path of a cross-sentence link. In this work, we make a naive assumption that two neighboring sentences share a “common root”. Therefore, a cross-sentence dependency path can be represented as two shortest dependency path branches from the ends to the “common root”, as shown in Figure 2.

Stanford CoreNLP[4] is used to parsing syntactic structures of sentences in this work.

### 3.2 Temporal Relation Classifiers

Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a natural choice for processing sequential dependency paths. As the reversed order also takes useful information, a backward representation can be achieved by feeding LSTM with the same input in reverse. We adopt the concatenation of the forward and backward LSTMs outputs, referred to as bidirectional LSTM (Graves and Schmidhuber, 2005).

Figure 3a shows the neural network architecture of our E-E, E-T classifier. Given an E-E or E-T temporal link, our system first generates two SDP branches: 1) the source entity to common root, 2) the target entity to common root. For each word along a SDP branch, concatenation of word, POS and dependency relation (DEP) embeddings (word-level) is fed into Bi-LSTM. The forward and backward outputs of both source and target branches are all concatenated, and fed into a fully connected hidden units layer. The final Softmax layer generates multi-class predictions. Since an E-D link contains single event SDP branch, our system applies a similar architecture, but with single branch Bi-LSTM with outputs fed into the penultimate hidden layer, as shown in Figure 3b.

In this work, we use word2vec[5] (Mikolov et al., 2013a,b) to train 200-dimensions word embeddings on English Gigaword 4th edition with skip-gram model and other default settings. For either of POS or DEP, we adopt the 50-dimensions lookup table initialized randomly.

| **LINK type** | **E-D** | **E-E** | **E-T** |
|---|---|---|---|
| AFTER | .493 | .477 | .350 |
| BEFORE | .552 | .380 | .311 |
| SIMULTANEOUS | - | - | - |
| INCLUDES | .305 | .185 | .254 |
| IS_INCLUDED | .513 | .296 | .204 |
| VAGUE | .482 | .656 | .616 |
| **Overall** | .491 | .544 | .480 |

Table 1: The best sentence-level 5-fold CV performance (Micro-average Overall F1-score).

## 4 Experiments

### 4.1 Hyper-parameters and Cross-validation

The grid search exploring a full hyper-parameter space takes time for three classifiers (E-E, E-T and E-D). Empirically, we set each single LSTM output with the same dimensions (equal to 300) as the concatenation of word, POS, DEP embeddings. The hidden layer is set as 200-dimensions.

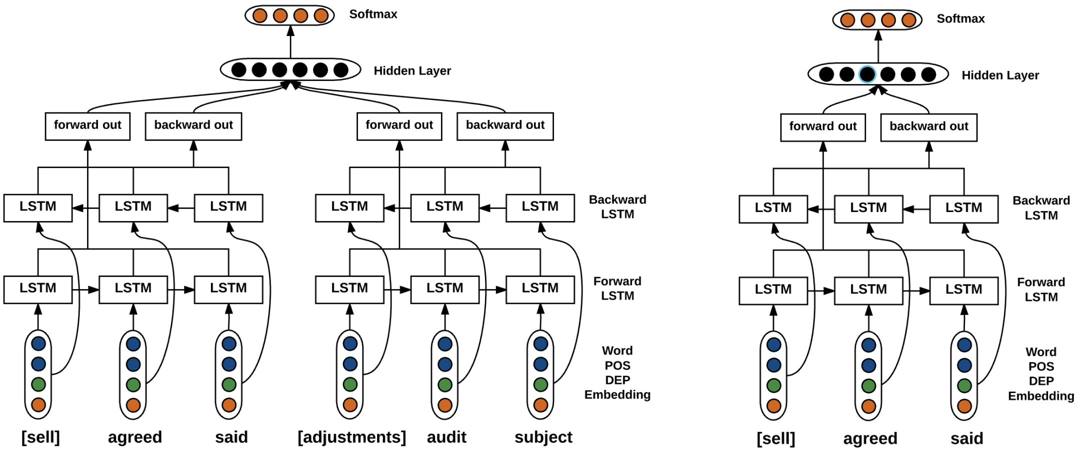
Our system adopts dependency paths as input, which means that the entities in the same sentences contain highly covered word sequence input. Simple cross-validation (CV) on links can not reflect the generalization ability of our model correctly. We use a grouped 5-fold CV based on the source entity ids (document id + sentence id) of links. This schema can reduce bias separately in either the source SDP or the target SDP. Although document level CV can avoid this issue, it’s not feasible for TimeBank-Dense because it contains only 27 training documents.

Early stopping is used to save the best model based on the validation data. In each run of the 5-fold cross-validation, we split 80% of ‘original training’ as ‘tentative training’ and 20% as ‘tentative test’. 85% of ‘tentative training’ is used to learning and 15% is used for validation. We also adopt early stopping in the final system on the validation data (15% of ‘original training’). The patience is set as 10.

Dropout (Srivastava et al., 2014) recently is proved to be an useful approach to prevent neural networks from over-fitting. We adopt dropout

[4] http://stanfordnlp.github.io/CoreNLP/

[5] https://code.google.com/archive/p/word2vec/



(a) E-E, E-T classifier

(b) E-D classifier

Figure 3: The DP-based Bi-LSTM temporal relation classifier.

| | Our | Mirza | Our | Mirza |
|---|---|---|---|---|
| LINK type | E-D | E-D | E-E | E-E |
| AFTER | **.582** | .466 | **.440** | .430 |
| BEFORE | .634 | **.671** | .460 | **.471** |
| SIMULTA. | - | - | - | - |
| INCLUDES | .056 | **.250** | .025 | **.049** |
| IS_INCLUD. | .595 | **.600** | .170 | **.250** |
| VAGUE | **.526** | .502 | **.624** | .613 |
| **Overall** | **.546** | .534 | **.529** | .519 |

Table 2: The detailed comparison of E-E and E-T against relation types to Mirza and Tonelli (2016) (Micro-average Overall F1-score) on test data.

separately after the following layers: embeddings, LSTM, and hidden layer to investigate the impact of dropout on performance. Table 1 shows the best CV results recorded in tuning dropout. The hyperparameter setting with the best CV performance is adopted in the final system.

### 4.2 Overall Performance

Recently, Mirza and Tonelli (2016) report state-of-the-art performance on TimeBank-Dense. They show the new attempt to mine the value of low-dimensions word embeddings by concatenating them with sparse traditional features. Their traditional features include entity attributes, temporal signals, semantic information of WordNet, etc., which means it's a hard setting for challenging their performance. In Table 2 and 3, 'Mirza' denotes their system.

| Systems | E-D | E-E | E-T | Overall |
|---|---|---|---|---|
| Baseline | .471 | .502 | .437 | .486 |
| Proposed | .546 | **.529** | .471 | **.520** |
| Mirza | .534 | .519 | .468 | .512 |
| CAEVO | **.553** | .494 | **.494** | .502 |

Table 3: The final comparison of E-E, E-T and E-D to the baseline and two state-of-the-art systems on test data.

Table 2 shows the detailed comparison to their work. Our system achieves higher performance on 'AFTER', 'VAGUE', while lower on 'BEFORE', 'INCLUDES' (5% of all data) and 'IS_INCLUDED' (4% of all data). It is likely that their rich traditional features help the classifiers to capture more minority-class links. On the whole, our system reaches better 'Overall' on both E-E and E-D. As their E-T classifier does not include word embeddings, the E-T results are not listed.

The final comparison is shown in Table 3. An one-layer fully connected hidden units baseline (200-dimensions) with word, POS embeddings as input (without any dependency information) is provided. The significant out-performance of our proposed model over the baseline indicates the effectiveness of the dependency path information and our Bi-LSTM in classifying temporal links. As a hybrid system, 'CAEVO' (Chambers et al., 2014) includes hand-crafted rules for their E-T and E-D classifiers. For instance, the temporal prepositions *in*, *on*, *over*, *during*, and *within* indicate 'IN_INCLUDED' relations. Their system is superior in E-T and E-D. 'Miza' takes the pure feature-based methods and performs slightly better in E-E and overall, compared to 'CAEVO'. Our system shows the highest scores in E-E and overall among the four systems. In general, our system achieves comparable performance to two state-of-the-art systems, without using any hand-crafted features, rules, or external resources.

## 5 Conclusion

We borrow the idea of the dependency path based neural networks into temporal relation classification. A "common root" assumption adapts our model to cross-sentence links. Our model adopts bidirectional LSTM for capturing both forward and backward orders information. We observe the significant benefit of the DP-based Bi-LSTM model by comparing it to the baseline. Our model achieves comparable performance to two state-of-the-art systems without using any explicit features (class, tense, polarity, etc.) or external resources, which indicates that our model can capture such information automatically.

## 6 Acknowledgments

We thank the anonymous reviewers for the insightful comments.

## References

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Vancouver, British Columbia, Canada, pages 724–731. http://www.aclweb.org/anthology/H/H05/H05-1091.

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)*. Association for Computational Linguistics, Baltimore, Maryland, pages 501–506. http://www.aclweb.org/anthology/P14-2082.

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. http://aclweb.org/anthology/Q/Q14/Q14-1022.pdf.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*. Association for Computational Linguistics, Barcelona, Spain, pages 33–40. https://www.aclweb.org/anthology/W/W04/W04-3205.pdf.

Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 918–927. http://www.aclweb.org/anthology/N13-1112.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(5):602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Natsuda Laokulrat, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. Uttime: Temporal relation classification using deep syntactic features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*. volume 2, pages 88–92. http://aclweb.org/anthology/S/S13/S13-2015.pdf.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, pages 753–760. https://doi.org/10.3115/1220175.1220270.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 368–375. https://doi.org/10.3115/1073083.1073145.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* http://arxiv.org/pdf/1301.3781v3.pdf.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119. https://arxiv.org/pdf/1310.4546.pdf.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

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*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2818–2828. http://aclweb.org/anthology/C16-1265.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1498–1507. http://www.aclweb.org/anthology/P13-1147.

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. https://catalog.ldc.upenn.edu/LDC2006T08.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 151–161. http://www.aclweb.org/anthology/D11-1014.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.

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* http://aclweb.org/anthology/S/S13/S13-2001.pdf.

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. https://doi.org/10.1007/s10579-009-9086-z.

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*. Association for Computational Linguistics, pages 57–62. http://www.aclweb.org/anthology/S10-1010.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015a. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 536–540. http://aclweb.org/anthology/D15-1062.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015b. 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*. Association for Computational Linguistics, Lisbon, Portugal, pages 1785–1794. http://aclweb.org/anthology/D15-1206.