Attention Neural Model for Temporal Relation Extraction

Sijia Liu[1,2], Liwei Wang[1], Vipin Chaudhary[2], Hongfang Liu[1]
[1]Department of Health Sciences Research, Mayo Clinic
{lastname.firstname}@mayo.edu
[2]Department of Computer Science and Engineering, University at Buffalo
vipin@buffalo.edu

Abstract

Neural network models have shown promise in the temporal relation extraction task. In this paper, we present the attention based neural network model to extract the containment relations within sentences from clinical narratives. The attention mechanism used on top of GRU model outperforms the existing state-of-the-art neural network models on THYME corpus in intra-sentence temporal relation extraction.

1 Introduction

A well-known challenge in leveraging electronic health records (EHRs) for research is to extract the information embedded in clinical texts. The recent progress in Natural Language Processing (NLP) techniques has facilitated the use of information in text for various clinical applications (Wang et al., 2017). One important NLP task in the clinical domain is to extract temporal relations between events and time expressions from clinical text for various EHR-based applications, such as clinical decision support and predictive modeling.

Along with studies in modeling clinical temporal events using structured EHR data (Zhao et al., 2017; Che et al., 2018), a series of temporal information extraction share tasks have been organized to encourage community efforts on the temporal relation extraction on unstructured clinical texts from EHR, such as i2b2 (Informatics for Integrating Biology and the Bedside) 2012 challenge (Sun et al., 2013) and Clinical TempEval shared tasks (Bethard et al., 2014, 2015, 2016). While both corpora are based on de-identified clinical notes, the major differences between i2b2 and TempEval are the evaluation and temporal event modeling. The i2b2 challenge evaluation enumerates all possible entity pairs from a clinical document into the evaluation, while the TempEval tasks leverage the concept of narrative containers which will enhance conventional temporal relations. In this study, we focus on the containment information extraction in TempEval.

In addition to the feature-based machine learning approaches such as Support Vector Machines (SVM) and conditional random field from top-performing TempEval 2016 systems (Lee et al., 2016; Abdulsalam et al., 2016; Tourille et al., 2016), there are several machine learning systems proposed after the shared task. Leeuwenberg and Moens (2017) used a structured learning method to predict temporal relations: Dligach et al. (2017) proposed an XML tag representation neural models such as Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) to mark the positions of the entities and achieved better performance compared to token position embeddings. They also evaluated the contains relations solely on medical events. Lin et al. (2016) experimented on different representations of XML tags proposed in (Dligach et al., 2017), and the results indicated that the input representation is an importance factor for the performance of neural models. A bidirectional LSTM (BiLSTM) approach has also been proposed in (Tourille et al., 2017). Their model utilized character embeddings to create a hierarchical LSTM model with corpus entities attributes as input into the embedding layer of their neural architecture. Recent related works using self-training (Lin et al., 2018) and human-like temporal reasoning via tree-based LSTM-RNN (Galvan et al., 2018) also achieved good performance in various evaluation scenarios, but direct comparisons are challenging due to differences in evaluation.

Inspired by visual attention models for object recognition in computer vision (Xu et al., 2015; Mnih et al., 2014), attention mechanism has also been successfully applied in several NLP tasks

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 134–139
Minneapolis, Minnesota, June 7, 2019. ©2019 Association for Computational Linguistics

such as machine translation (Luong et al., 2015), machine reading (Cheng et al., 2016), document classification (Yang et al., 2016) and relation extraction (Lin et al., 2016), to obtain state-of-the-art performance. The attention layer oversees the entire sequence of recurrent neural network (RNN) units and is trained to pay more "attention" to salient units.

In this paper, we present an attention neural model to identify containment relations from clinical narratives with annotated medical events and temporal information. The model achieves state-of-the-art performance in intra-sentence temporal relation extraction while using minimal entity features and external knowledge.

2 Materials

We use the THYME (Temporal Histories of Your Medical Event) corpus (Styler IV et al., 2014) to evaluate our proposed models. THYME corpus is extracted from Mayo Clinic colon cancer data, which contains clinical notes from 200 patients. The corpus is manually de-identified to remove patient identification, and is fully annotated into two types of entities: Timex3 and Event. Timex3 contains temporal information like event dates and timestamps. The definition of event is a broad concept of patient health related conditions and mentions.

All the Event entities contain 5 attributes, "Modality", "Degree", "Polarity", "Type" and "DocTimeRel". The Document Time Relations (DocTimeRel) specifies the temporal relation of the event to the time of service. In this study, we focused on the temporal relations between two different entities within one sentence, namely intra-sentence relations as referred in (Tourille et al., 2017). Therefore, we did not include DocTimeRel, which is an event attribute, into our model and evaluation.

3 Methods

We define the temporal relation extraction problem as a relation classification problem among relation candidates generated from annotated entities. Specifically, for all the events within one sentence, we enumerate all possible entity pairs as relation candidates. Then, we assign relation labels based on the gold standard annotations provided with the corpora. In THYME corpus, the gold standard annotations consist of relation between two entities and its relation type. When we prepare the dataset for relation classification, for each combination of entities, we have three potential labels: 1) the first entity "CONTAINS" the second entity in temporal; 2) the first entity is "CONTAINED" by the second entity; 3) the two entities do not have a containment temporal relation, i.e. "NA".

3.1 Input Representation

Given clinical narratives with annotated entities, we first use the Punkt sentence tokenizer[1] to separate the sectionized raw text into section titles and sentences. Then an associated encoding of entities into XML tags are constructed, following the work of Lin et al (Lin et al., 2017). The event entities are surrounded by "<e>" and "</e>". The temporal entities are replaced by the special XML tags from time class provided with the entity annotations, e.g. "<time>", "<date>", "<duration>" and "<prepostexp>", and surrounded by "<t>" and "</t>". In our preliminary experiments, this entity representation also leads to better results than position embeddings, which use relative distances between two entities as index to compute the high-dimensional embeddings of each word (Zeng et al., 2014).

3.2 Attention Neural Models

To improve the system performance of neural network models, we would like to leverage the emerging attention mechanism. Attention based RNN uses an attention layer to capture the salient units of a sequence by maintaining a context vector for the sequence models. Word-level attention weights can be interpreted as importance measure in given contexts, i.e. temporal relation indicators for each relation instance of a sentence. The architecture of our proposed model is shown in Figure 1. In the example, the entities "monitored" and "three months" are surrounded by the XML tags introduced above. The Timex3 entity "three months" is replaced by the entity type "<duration>" when feeding into the word embedding layer. Ideally, a high attention weight will be given to the preposition "in", as it is the word expressing the containment relations between the event and the time. Besides, the entity and the tags may also need to contribute to the discrimination of different relation types.

[1] https://www.nltk.org/_modules/nltk/tokenize/punkt.html

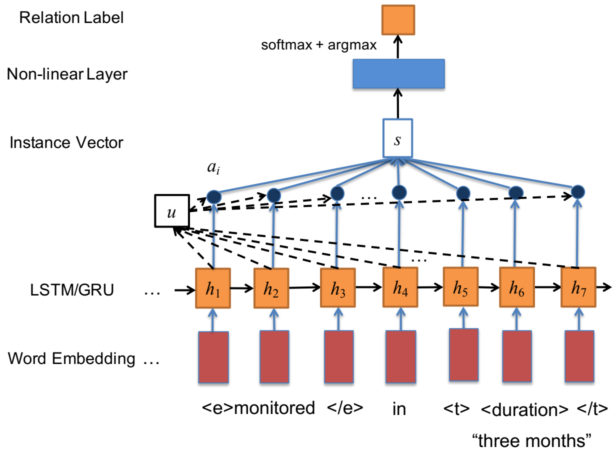

Figure 1: The architecture of attention based RNN for temporal relation extraction.

The vectors of RNN units are denoted as h_i, where i is the index of the input tokens in the generated relation instances. Similar to (Yang et al., 2016), we would like to obtain a word-level attention weights a_i for each entity pair, which is calculated based on the sequence of RNN outputs, either LSTM or Gated Recurrent Unit (GRU) proposed in (Cho et al., 2014). To reward the salient units for relation classification, a trainable context vector u_v is used to retrieve the attention weights a_i, and it is computed from trainable parameters W_v and b_v from the attention layer. The word-level attention weight a_i is calculated using a softmax function. Afterwards, the sentence vector s is computed as the weighted sum of a_i. Specifically, the sentence vector can be computed as:

$$u_i = \tanh(W_v h_i + b_v),$$

$$a_i = \frac{\exp(u_i^T u_v)}{\sum_i \exp(u_i^T u_v)},$$

$$s = \sum_i a_i h_i.$$

The word embedding, RNN and attention layers combined can be regarded as an instance encoder. For each relation instance generated as described in Section 3.1, those layers together encode the instance into a multi-dimensional vector s. The encoded relation instance vector s is then fed into a fully connected layer. The output dimension of the fully connected layer is set to the number of potential labels, which is 3 in this study.

Then, a softmax function normalizes the outputs into a predicted probability of 3 labels, where the sparse cross entropy loss is calculated and minimized during training. We take the maximum probability as the relation label for the evaluation of closure-enhanced precision, recall and F1-score.

3.3 Evaluation

The official evaluation scripts of TempEval[2] use the concept of "narrative containers" (Miller et al., 2013) to validate the results. Narrative containers is a set of events that contains multiple temporal relations. The official evaluation uses narrative container to evaluate the system performance, instead of evaluating directly from the relation classification results by instances. The usage of closure is intended to reduce the penalty caused by extracting the implicit relations that can be inferred between events but are not included in the annotation.

Following the shared task of TempEval 2016 and recent related work on the THYME corpus, we focus on the extraction of temporal containment relations. This is because the prevalence of contains relations is much higher than other temporal relations.

4 Experiments and Discussion

We ran our experiments on similar settings as (Dligach et al., 2017). The cross sentence relations are excluded in our evaluation.

The models are implemented in Keras with Tensorflow backend. The experiments are done on a computing server with NVIDIA Tesla P40 GPU. Each epoch of attention based LSTM took approximately 300 seconds while GRU will take approximately 250 seconds, due to fewer trainable parameters needed for each unit.

The 300-dimension word embeddings from Glove-6B[3] are selected as the input based on our preliminary experiments on trained embeddings from biomedical domain (Wang et al., 2018) as well as the THYME corpus. The embedding of out-of-vocabulary words, including special XML tags, are determined by random sampling from unit distribution in [-0.1, 0.1]. The hyperparameters are selected based on the optimal combination from the development set when training on

[2] https://github.com/bethard/anaforatools

[3] https://nlp.stanford.edu/projects/glove/

Model	Event-Time			Event-Event		
	P	R	F1	P	R	F1
THYME (Dligach et al., 2017)	0.577	0.845	0.685	0.595	0.572	**0.584**
CNN tokens (Dligach et al., 2017)	0.683	0.717	0.700	0.688	0.412	0.515
ATT-LSTM	0.770	0.722	0.744	0.535	0.582	0.558
ATT-GRU	0.765	0.737	**0.750**	0.617	0.550	0.579

Table 1: Performance comparison in Event-Time and Event-Event containment relations on test set

Model	P	R	F1
BiLSTM (Tourille et al., 2017)	0.670	0.681	0.675
BiLSTM + cTAKES (Tourille et al., 2017)	0.663	0.704	0.683
ATT-LSTM	0.687	0.666	0.676
ATT-GRU	0.698	0.684	**0.690**

Table 2: Performance comparison in intra-sentence containment relations on test set

the training set. To avoid potential overfitting during the training phase, we apply drop out technique (Srivastava et al., 2014) with the drop out rate of 0.5. Adam optimizer (Kingma and Ba, 2014) is used with learning rate 0.001 to train the model and sparse categorical cross entropy as the loss function. We apply early stopping during training to avoid overfitting by terminating the training process if there is no validation accuracy increase in consecutive 4 epochs. Then the training and development set are combined to train the model while tested on the testing set. The batch size of training is 64, and the unit size for RNN units is set 128 based on hyperparameter tuning.

The evaluation results on Event-Time and Event-Event relation extraction in closure-enhanced precision (P), recall (R) and F1-score are shown in Table 1. "ATT-" denotes our attention based RNN models. The results in Table 1 are directly comparable with the work in (Dligach et al., 2017), since the models of Event-Time Event-Event relations are trained separately. The most significant improvement is from the Event-Time relation extraction, where the ATT-GRU (0.750) outperforms the CNN model by 0.050. In the Event-Event relations, ATT-GRU model outperforms the CNN model, but is not as good as the feature based SVM model in the THYME system (-0.05). One potential reason for the performance gain is that the ATT models oversee all units from the RNN layer rather than focusing on the max pooling of local features as CNN.

When we combined both Event-Time and Event-Event relations together, Table 2 shows the results for all temporal relations within each sentence. Compared to other neural network models, our proposed ATT-GRU (0.690 F1) is favorably comparable to the BiLSTM model incorporating cTAKES outputs[4] (BiLSTM+cTAKES) and character embeddings (+0.007). We only use the raw text and annotated entity types, while BiLSTM+cTAKES requires finer granularity of the UMLS[5] entity types and semantic types as inputs. It is our future perspective to utilize character-level information and entity attributes as the input to further improve our system. ATT-GRU performs better than LSTM in all the three evaluation scenarios. One potential reason is that GRU has less trainable parameters compared to LSTM, thus it may converge better in a corpus with relatively limited positive relational instances.

One challenge for neural models in the temporal relation extraction task is class imbalance. The majority of the errors are caused by the confusion between negative ("NA") and positive ("CONTAINS"+"CONTAINED") instances, while very few of the errors are from the confusion of "CONTAINS" and "CONTAINED" relations. The ratios between positive and negative relations of Event-Event, Event-Time and those combined are 1:3.4, 1:12.7 and 1:8.4, respectively. The class weights are tuned in the feature-based THYME system to improve the balance of precision and recall, but there is no such effort on other neural models in both our work and (Dligach et al., 2017).

[4] https://ctakes.apache.org/

[5] Unified Medical Language System: https://www.nlm.nih.gov/research/umls/

Lin et al. (2017) analyzed the impact of different XML tags for the temporal entities as inputs. The one-token tag representation for multi-word temporal repressions (e.g. replacing the Timex3 mention “March 11, 2014” by “<date>”) shows improvements on the classification, which is also used in our study. Compared to Lin’s method, our model is a single neural model instead of a combined model of CNN and SVM for Event-Event and Event-Time relations, respectively. Leeuwenberg and Moens (Leeuwenberg and Moens, 2017) used structured learning on all relations within token distance of 30. The framework can also be extended to model inter-sentence relations by adding such relation instances into the training and testing, but fine-tuned down-sampling needs to be done to optimize its performance.

5 Conclusion and Future Work

In this paper, we presented the attention-based neural networks on temporal relation extraction. The proposed attention based GRU model achieved state-of-the-art performance in intra-sentence containment temporal relation extraction on THYME corpus.

In future, we would like to adopt the hierarchical model with character embeddings in the word-level representation into our attention based neural networks. We would also like to explore the comparison between different variations of the attention mechanisms such as multi-head attention (Vaswani et al., 2017) and self-attention (Cheng et al., 2016; Verga et al., 2018).

Acknowledgment

We thank anonymous reviewers for their constructive feedbacks to this manuscript. The work is supported by National Library of Medicine grant R01LM011829 and National Science Foundation IPA grant. Opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health or National Science Foundation.

References

Abdulrahman AAl Abdulsalam, Sumithra Velupillai, and Stephane Meystre. 2016. Utahbmi at semeval-2016 task 12: Extracting temporal information from clinical text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1256–1262.

Steven Bethard, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2014. Clinical tempeval. *arXiv preprint arXiv:1403.4928*.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *SemEval@ NAACL-HLT*, pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dmitriy Dligach, Timothy A. 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 746–751. 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. Association for Computational Linguistics.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 1292–1297. The Association for Computer Linguistics.

Tuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Chen Lin, T. Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP*.

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. Association for Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *ACL*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Timothy A Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana K Savova. 2013. Discovering narrative containers in clinical text. *ACL 2013*, page 18.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212.

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.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. Limsi-cot at semeval-2016 task 12: Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1136–1142.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 224–230. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. 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 872–884. Association for Computational Linguistics.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2017. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.

Jing Zhao, Panagiotis Papapetrou, Lars Asker, and Henrik Boström. 2017. Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics*, 65:105–119.