End-to-end clinical temporal information extraction with multi-head attention

Timothy Miller
Computational Health Informatics Program
Boston Children's Hospital
Harvard Medical School
timothy.miller@childrens.harvard.edu

Steven Bethard
School of Information
University of Arizona
bethard@arizona.edu

Dmitriy Dligach
Department of Computer Science
Loyola University, Chicago
ddligach@luc.edu

Guergana Savova
Computational Health Informatics Program
Boston Children's Hospital
Harvard Medical School
guergana.savova@childrens.harvard.edu

Abstract

Understanding temporal relationships in text from electronic health records can be valuable for many important downstream clinical applications. Since Clinical TempEval 2017, there has been little work on end-to-end systems for temporal relation extraction, with most work focused on the setting where gold standard events and time expressions are given. In this work, we make use of a novel multi-headed attention mechanism on top of a pre-trained transformer encoder to allow the learning process to attend to multiple aspects of the contextualized embeddings. Our system achieves state of the art results on the THYME corpus by a wide margin, in both the in-domain and cross-domain settings.

1 Introduction

Temporal information extraction is the task of discovering event mentions, time expressions, and relations between them that indicate their relative temporal ordering. The goal of these tasks is to place events on a timeline with as specific granularity as possible, to better understand topics that have strong temporal components. In the clinical setting, temporal information extraction can be applied to the text in electronic health records to create a timeline for a patient, enabling downstream applications that are heavily time-dependent (e.g., modeling the temporal ordering of problems and medication use to extract causally plausible candidates for adverse drug events). Recent datasets have been released as part of community shared tasks that enable the research community to make progress on this challenging suite of tasks (Styler IV et al., 2014; Sun et al., 2013a; Wright-Bettner et al., 2019; Bethard et al., 2015, 2016, 2017). The THYME (1 and 2) and i2b2 datasets have led to progress in this area but these problems are far from solved.

Both the Clinical TempEvals (Bethard et al., 2015, 2016, 2017) and the 2012 i2b2 Challenge (Sun et al., 2013b) evaluated systems on an end-to-end basis – that is, given only raw text, participants needed to both extract relation arguments and relations between them. However, most systems attacked this setting with a pipeline approach, combining the best event and time extraction systems with a relation system. In recent general domain work, models for doing end-to-end relation extraction have been obtaining some success recently by building on pre-trained transformers in the BERT family (Devlin et al., 2019), using additional task-specific components that aggregate encoder information into relation predictions (Liu et al., 2020; Sui et al., 2020).

In this work, we extend Liu et al. (2020)'s model for the task of clinical temporal information extraction. Instead of using multi-headed attention for different relation categories, we use multiple heads to capture different aspects of the token representations. The additional layer allows the model to then integrate the signals to make the final relation predictions.

Our results show that this architecture can obtain state of the art performance for relation extraction on the Clinical Tempeval 2016 and 2017 tasks, which evaluated in-domain (colon cancer) and cross-domain (colon cancer → brain cancer) settings, respectively. Our analysis shows that, for this task, multi-headed attention is far superior to

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 313–319

single-headed attention with the same number of model parameters, suggesting that the multi-headed approach is valuable for representing different aspects of the input. In addition, we compare the trade-offs in modeling this as a multi-task learning setup, as opposed to using multiple fine-tuned classifiers that work separately, where the time expressions and events are found separately and used as the anchors for the relation classifier. Finally, we report results of this model on the THYME2 data, for which no current baseline results exist in the end-to-end task.

2 Background and Related Work

This work takes advantage of the THYME (Temporal History of Your Medical Events) corpus (Styler IV et al., 2014), and specifically the data released as part of the Clinical Tempeval shared tasks hosted at SemEval (Bethard et al., 2015, 2016, 2017). This English-language dataset uses colorectal and brain cancer notes from patients at the Mayo Clinic. Each patient is represented with 3-4 notes written around the time of cancer diagnosis, typically with two clinical notes and one additional radiology and/or pathology note. The first two Clinical TempEval's evaluated on colon cancer only.

The third Clinical TempEval was a domain adaptation task where participants were given the colon cancer data and evaluated on the brain cancer test set. There were two settings, one with unsupervised domain adaptation (no access to brain training data), and another with a small amount of labeled brain cancer data. In practice, there was little unsupervised adaptation attempted, with one top system obtaining negative results with the technique of freezing embedding weights before training (to prevent some of the embeddings from updating and drifting away from target-domain-specific terms that might not be seen during training) (Tourille et al., 2017), and another also obtaining negative results by replacing rare terms in the training and test data with an UNK token (Leeuwenberg and Moens, 2017). In this work we use the THYME colon cancer data for training, the development set for tuning and model selection, and evaluate on both the colon cancer test set and brain cancer test set, in the unsupervised setting.

More recent work on THYME has focused on the easier relation *classification* setting, where gold standard time expressions and events are provided, and the system task is to decide which pairs of temporal entities should be linked with temporal relations. The current best-performing system is from Lin et al. (2021). That paper's main contribution was a continued pre-training technique for clinical text, where the masked language modeling task was modified to preferentially mask tokens that were part of temporal entities. The temporal relation extraction was modeled as a sentence classification task, where the gold standard arguments were marked by special tokens in the input, and a softmax layer off of the pre-trained transformer's sentence representation token was used to predict the relation between the two marked tokens (including the None relation). Such an approach cannot be fairly compared (or even deployed) without building additional systems to do the event and time expression detection, so we do not directly compare to that result.

Other work in end-to-end relation extraction is also relevant to this work. The most relevant related work (Liu et al., 2020) used a BiLSTM encoder to get token representations, which fed into a multi-headed attention layer, where each attention head was used to score a different relation type. That work differs from ours in that it is applied only to "general-domain" tasks including the NYT (Riedel et al., 2010) and WebNLG datasets (Gardent et al., 2017), and treats outputs values from different attention heads as prediction scores for individual relation categories. Our work, in contrast, uses attention scores as features for a single downstream relation classifier (the CONTAINS relation), allowing the model to use different aspects of the representation to make classification decisions.

The current work also has connections to techniques for end-to-end relation extraction where the task was treated as a table filling task (Gupta et al., 2016), where each cell (i, j) in the table represents the relation between token i and token j.

3 Methods

Our model uses a multi-task learning setup built on top of a pre-trained transformer encoder, and specifically the PubmedBERT-base-uncased-abstract model (Gu et al., 2021). Time expressions and events are modeled as two independent Begin-Inside-Outside (BIO) tagging tasks, so that each contextualized token embedding in the sequence is used to make a classification decision for that token's categorization as an event or time expression. This BIO tagging is implemented with softmax

layer at every token index over the B- or I- version of each of the time expression or event types. Relation classification is treated as a binary task applied to n^2 token pairs to find relations, using a mechanism very similar to the self-attention mechanism already used extensively in transformers. Our attention-inspired classification model is an additional layer on top of the pre-trained transformer that takes in the contextualized token representations from the last layer of the transformer encoder.

Attention first came to wide use in encoder-decoder architectures (e.g. (Bahdanau et al., 2015; Cho et al., 2015)), as a way for a decoder to pick out relevant sub-components of the input (e.g., individual token representations for indeterminately long input sentences) during generation of outputs. Typically, it is implemented as a fully-connected neural layer that learns a score for each sub-component (Kim et al., 2017) and uses the computed weighted average over the sub-components at each decoding step. Self-attention is a simple extension where the attention mechanism is used to update weights over a discrete set of elements at a higher layer l by applying attention over the same elements at a lower layer $l-1$. Vaswani et al. (2017), in introducing the transformer architecture, showed that attention alone could be used both in encoding and decoding to learn hidden representations. Further, they introduced the idea of *multi-headed attention*, that is, the idea that several different attention distributions can be computed, as a way to focus on and combine multiple different aspects of the input representations.

The first stage in attention is projecting the input with Query, Key, and Value matrices, and then the attention is computed as:

$$Attention(Q,K,V) = softmax\frac{(QK^T)}{\sqrt{d_k}}V$$

Multi-headed attention multiplies this out with distinct Q, K, and V projections for H attention heads (indexed by h), and then concatenates the outputs, with the sizes of the projection matrices scaled such that the total output dimensionality matches the input. For self-attention, the inputs to the Q, K, and V projections are all the same – the output from the previous layer.

The most important aspect of attention for our purposes is that the matrix product $S^h = Q^h(K^h)^T$ in the numerator results in a matrix of size $N \times N$ for N tokens. In the work by Liu et al. (2020), each relation type is aligned to one attention head, so the value at each position in the matrix represents the affinity score for the relation aligned to head h:

$$s^h(i,j) = S^h(i,j)$$

In this work, we use H attention matrices to compute a single $H \times N \times N$ tensor, where a softmax along the attention head dimension leads to a classification for each token pair into a single relation, the temporal narrative container relation:

$$p_c(i,j) = softmax([S^0_{i,j}, S^1_{i,j}, \cdots S^H_{i,j}])$$

The overall loss function combines time expression extraction, event extraction, and relation classification, all of which use versions of cross-entropy loss. The overall loss function simply sums the three, with a weight on the relation loss that we tune on the development set to maximize F1 score for the relation classification task.

$$L_{total} = L_{timex} + L_{event} + \alpha_{rel} * L_{rel}$$

4 Evaluation

We evaluate on the THYME colon and brain corpora, tuning hyperparameters on the colon dev set and reporting results of the narrative container extraction task on the colon and brain test sets.[1] Hyperparameters we tuned over included fine tuning learning rate, batch size, classifier layer (which layer the classification head is connected to), relation task weight, number of classifier attention heads, and dimensionality of attention heads. For all reported results we use the official SemEval scoring tool, which performs transitive closure for the relations (UzZaman and Allen, 2011). Table 1 shows the primary outcomes, the precision, recall, and F1 scores for the narrative container relation. We compare against the best-performing system from the relevant Clinical TempEval shared task – for colon cancer it was a system from UTHealth (Lee et al., 2016) and for brain cancer it was a system from GUIR (MacAvaney et al., 2017).

While the system we describe is multitask in nature, it does not have explicit interactions between the sequence tagging models and the relation predictor. We first find predicted relations between tokens and work backwards to find the argument span

[1] SemEval 2016 and 2017 also used only the narrative container extraction task due to its superior inter-annotator agreement.

that exists at each token to make a relation prediction. This independence allows us to investigate the importance of the quality of the extraction of those entities, by using alternative methods, including an oracle, in the second step of finding the arguments for predicted relations. The row labeled + *Gold args* shows the performance with oracle-based arguments taken from the gold standard, to show us an upper bound on how well relation extraction would work with perfect argument extraction. The row labeled - *MTL* shows the performance if we use separately tuned event and time expression extraction systems, since multi-task learning sometimes can cause degraded performance of independent components.

The E2E-MHA system obtains the best performance on both corpora. For the colon cancer corpus, precision is 14 points higher than the best performing Clinical TempEval system, while recall is within one point, giving a 5-point increase in F1. For the brain cancer corpus, precision decreases by 11 but the increase in recall results in a small 2-point increase in F1.

The next result we report is on the importance of multi-headed attention. Our hypothesis is that using multiple attention heads allows the model to more easily learn to focus on multiple facets of an instance. An alternative hypothesis is that the gain comes from using a lot more parameters than an architecture that uses one attention head per relation (e.g., (Liu et al., 2020)). To test this, we can compare against a similar model that has fewer attention heads but more feature dimensions per head, holding the total number of model parameters constant. For this experiment, we report colon development set validation scores directly as measured by scikit-learn (i.e., not using Clinical TempEval scorer, so it does not include temporal closure), as we tuned this parameter before running final tests.

Table 2 shows the results of this experiment. We found there was a sweet spot with 256 attention heads of size 64 each, where a similarly parameterized versions lose performance. As the number of heads decreases to $<= 4$, performance drops to an extent that finding good hyperparameter settings was not possible in our limited search.

Finally, we report the results when the same system is tuned, trained, and applied to the THYME2 corpus (Wright-Bettner et al., 2020). THYME2 introduced a new relation type called NOTEDON which was used to indicate a test noting a finding, which had previously been covered by CONTAINS. This change improved annotator agreement and removed logical inconsistencies that were created by the previous schema. Another new relation type in THYME2 is CONTAINS-SUBEVENT, which was primarily used for cross-document annotations. For THYME2 we report results micro-averaged across all relations in Table 1, and break these results down by relation category in Appendix A. As far as we are aware, this is the first reported end-to-end temporal relation extraction result on THYME2.

Test Split	System	Prec	Rec	F1
Colon	UTHealth1	0.49	**0.47**	0.48
	E2E-MHA	**0.63**	0.46	**0.53**
	+ Gold args	0.74	0.49	0.59
	- MTL	0.66	0.44	0.53
Brain	GUIR	**0.52**	0.25	0.34
	E2E-MHA	0.41	**0.32**	**0.36**
	+ Gold args	0.64	0.41	0.50
	- MTL	0.42	0.25	0.32
THYME2	E2E-MHA	0.65	0.46	0.54

Table 1: Results of the end-to-end multi-headed attention system (*E2E-MHA*) on the end-to-end narrative container relation extraction task. Best non-oracle results shown in **bold**. THYME2 is colon cancer only, and is a micro-average across relation types, while THYME1 results are on the narrative container (CONTAINS) relation only, following SemEval, so the results are not directly comparable.

Attn heads	Head size	F1
256	64	0.58
64	256	0.57
16	1024	0.57

Table 2: Narrative container extraction performance with different numbers and sizes of attention heads. The top line is the system used in the previous experiments.

5 Discussion and Conclusion

While this task remains challenging, with the highest obtained F1 scores 0.53 (in-domain) and 0.36 (out-of-domain), the combination of biomedically pre-trained transformers and a novel multi-headed attention classifier obtains results that improve the state of the art. Even in-domain, the task is fairly far from being solved.

Additional future work should also explore explicit domain adaptation, as out-of-domain perfor-

mance still suffers despite obtaining better performance than previous methods. Simple unsupervised methods like target domain continued pretraining (Gururangan et al., 2020) are likely to provide benefits on the brain cancer corpus, and future work should investigate this and other approaches. We did see performance improvements in preliminary work when switching from general domain pre-trained models (roberta-base Liu et al. (2019)) to a biomedically trained model, so it is possible that some of those potential gains are already incorporated by the more general language model.

The trained models developed for this work are available in the HuggingFace Models repository.[2] The code for implementing and evaluating these models is available on GitHub.[3] The code repository also contains a FastAPI-based demonstration server that sets up a REST server to convert text arguments into temporal relations in a simple JSON schema.

Acknowledgements

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Numbers R01LM012973, R01LM012918, and R01LM010090, by the National Institue on Drug Abuse under Award Number R01DA051464, and by the National Institute Of Mental Health of the National Institutes of Health under Award Number R01MH126977. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

S Bethard, L Derczynski, and G Savova. 2015. Semeval-2015 task 6: Clinical tempeval. *Proc. SemEval*.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Conference on Semantic Evaluation (SemEval 2016)*.

[2]https://huggingface.co/mlml-chip

[3]https://github.com/Machine-Learning-for-Medical-Language/cnlp_transformers

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2537–2547, Osaka, Japan. The COLING 2016 Organizing Committee.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

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-2016)*, pages 1292–1297, San Diego, California. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. 2017. KULeuven-LIIR at SemEval-2017 task 12: Cross-domain temporal information extraction from clinical records. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1030–1034, Vancouver, Canada. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

Jie Liu, Shaowei Chen, Bingquan Wang, Jiaxin Zhang, Na Li, and Tong Xu. 2020. Attention as Relation: Learning Supervised Multi-head Self-Attention for Relation Extraction. In *IJCAI*, pages 3787–3793.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

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.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2020. Joint Entity and Relation Extraction with Set Prediction Networks.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46 Suppl:S5–12.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–13.

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 597–602, Vancouver, Canada. Association for Computational Linguistics.

Naushad UzZaman and James F Allen. 2011. Temporal Evaluation. *The 49th Annual Meeting of the Association for Computational Linguistics Human Language Technologies*, 271(5):351–356.

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*, volume 30. Curran Associates, Inc.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H. Martin, and Guergana Savova. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114, Online. Association for Computational Linguistics.

Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller. 2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong. Association for Computational Linguistics.

A THYME2 Results by Relation Category

Relation type	Prec	Rec	F1
BEFORE	0.468	0.262	0.336
BEGINS-ON	0.606	0.275	0.378
CONTAINS	0.699	0.669	0.684
ConSub	0.369	0.044	0.079
ENDS-ON	0.641	0.210	0.316
NOTED-ON	0.721	0.652	0.685
OVERLAP	0.513	0.231	0.318
Micro-F	0.654	0.461	0.541

Table 3: Results of the end-to-end system trained and tested on THYME2. The new NOTED-ON category has excellent performance, while CONTAINS maintains strong performance. This is not surprising as these are the categories with the most data. CONTAINS-SUBEVENT (ConSub) performs the worst, but this is expected as it is a cross-document relation and this system is optimized for a local token context.