A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction

Chen Lin[1], Timothy Miller[1], Dmitriy Dligach[2], Steven Bethard[3] and Guergana Savova[1]

[1]Boston Children's Hospital and Harvard Medical School
[2]Loyola University Chicago
[3]University of Arizona
[1]{first.last}@childrens.harvard.edu
[2]ddligach@luc.edu
[3]bethard@email.arizona.edu

Abstract

Classic methods for clinical temporal relation extraction focus on relational candidates within a sentence. On the other hand, breakthrough Bidirectional Encoder Representations from Transformers (BERT) are trained on large quantities of arbitrary spans of contiguous text instead of sentences. In this study, we aim to build a sentence-agnostic framework for the task of CONTAINS temporal relation extraction. We establish a new state-of-the-art result for the task, 0.684F for in-domain (0.055-point improvement) and 0.565F for cross-domain (0.018-point improvement), by fine-tuning BERT and pre-training domain-specific BERT models on sentence-agnostic temporal relation instances with WordPiece-compatible encodings, and augmenting the labeled data with automatically generated "silver" instances.

1 Introduction

The release of BERT (Devlin et al., 2018) has substantially advanced the state-of-the-art in several sentence-level, inter-sentence-level, and token-level tasks. BERT is trained on very large unlabeled corpora to achieve good generalizability. Instead of relying on a recurrent neural network, BERT uses a transformer architecture to better capture long distance dependencies. BERT is able to make predictions that go beyond natural sentence boundaries, because it is trained on fragments of contiguous text that typically span multiple sentences.

These advantages of BERT motivate us to apply it to a traditionally sentence-level task – temporal relation extraction from clinical text. The identification of temporal relations in the clinical narrative can lead to accurate fine-grained analyses of many medical phenomena (e.g., disease progression, longitudinal effects of medications), with a variety of clinical applications such as question answering (Das and Musen, 1995; Kahn et al., 1990), clinical outcomes prediction (Schmidt et al., 2005), and recognition of temporal patterns and timelines (Zhou and Hripcsak, 2007; Lin et al., 2014). However, the labeled instances for this clinical information extraction task are limited, so neural models trained from scratch may not be able to learn complex linguistic phenomena. Pre-trained models like BERT could potentially provide rich representations as they are trained on massive data.

Classic models for clinical temporal relation extraction have framed the task within a sentence (Sun et al., 2013; Bethard et al., 2015, 2016, 2017), making them susceptible to sentence detection errors. Using BERT, on the other hand, eliminates this sensitivity to sentence boundary errors. The key contributions of this paper are: (1) introducing BERT to the challenging task of clinical temporal relation extraction and evaluating its performance on a widely used testbed (THYME corpus; Styler IV et al., 2014), (2) developing a universal processing mechanism based on a fixed, sentence-boundary agnostic window of contiguous tokens, (3) pre-training BERT on MIMIC-III (Medical Information Mart for Intensive Care) dataset (Johnson et al., 2016) and comparing its performance to BERT and its biomedical adaptation BioBERT (Lee et al., 2019), (4) augmenting the labeled set with automatically generated instances from unlabeled data, and (5) evaluating models for in- and cross-domain tasks on the THYME corpus.

2 Background

Recently, several pre-trained general-purposed language encoders have been proposed, including CoVe (McCann et al., 2017), ELMo (Peters et al., 2018), Flair (Akbik et al., 2018), GPT (Radford et al., 2018), GPT2 (Radford et al., 2019), and BERT (Devlin et al., 2018). These models are trained on vast amounts of unlabeled text to achieve

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 65–71
Minneapolis, Minnesota, June 7, 2019.

generalizable contextualized word embeddings, and some can be fine-tuned to fit a supervised task.

BERT is trained using a masked language model and the next-sentence objectives. Its architecture consists of stacked multi-layered transformers, each implementing a self-attention mechanism with multiple attention heads. BERT can be further pre-trained for specific domains (Lee et al., 2019) or serve as a backbone model to be fine-tuned with one output layer for a wide range of tasks.

For the task of clinical temporal relation extraction, recent years have seen the rise of neural approaches – structured perceptrons (Leeuwenberg and Moens, 2017), convolutional neural networks (CNNs) (Dligach et al., 2017; Lin et al., 2017), and Long Short-Term memory (LSTM) networks (Tourille et al., 2017; Dligach et al., 2017; Lin et al., 2018) – where minimally-engineered inputs have been adopted over heavily feature-engineered techniques (Sun et al., 2013). The THYME corpus (Styler IV et al., 2014), which is annotated with time expressions (TIMEX3), events (EVENT), and temporal relations (TLINK) using an extension of TimeML (Pustejovsky et al., 2003; Pustejovsky and Stubbs, 2011), is a popular choice for evaluation and was used in the Clinical TempEval series (Bethard et al., 2015, 2016, 2017).

CONTAINS relations are by far the most frequent type of relation in the THYME corpus. They signal that an EVENT occurs entirely within the temporal bounds of a *narrative container* (Pustejovsky and Stubbs, 2011). The THYME corpus is limited in size so models developed on it may suffer from low generalizability. Recent efforts to improve performance have attempted tree-structured models (Galvan et al., 2018) or assistance from unlabeled data (Lin et al., 2018). Years of shared work on this problem and plateauing scores may have suggested that performance on this task is at its peak. However, given the successful application of BERT on many different tasks in the general domain, as well as more recent work in relation extraction tasks (Wang et al., 2019; Lee et al., 2019), we wanted to explore applying this new model to the clinical temporal relation extraction task.

Conventionally, the tasks of within- and cross-sentence relation extraction have been treated separately (Sun et al., 2013; Tourille et al., 2017) as they call for different features. While some methods focus on within-sentence relations (as they are the majority), such methods are susceptible to sentence-boundary detection errors. The input sequences of arbitrary lengths that BERT operates on cover both within-sentence and cross-sentence situations, enabling us to design a universal model that is sentence boundary agnostic.

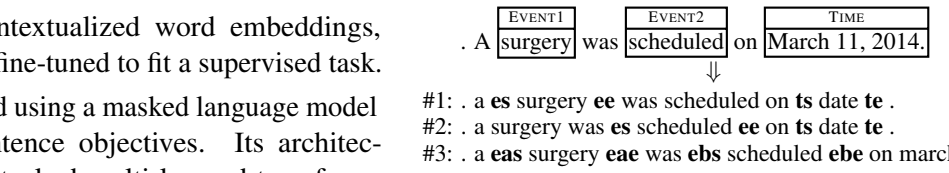

Figure 1: Representations of three candidate relations produced from an example token sequence.

3 Methods

3.1 Task definition

We process the THYME corpus using the segmentation and tokenization modules of Apache cTAKES (http://ctakes.apache.org). We consume gold standard event annotations, gold time expressions and their classes (Styler IV et al., 2014) for generating instances of containment relation candidates. Each instance consists of a pair of event entities, or an event entity and a time expression entity. We preserve the natural order of the two entities in their original context and represent the instance as a sequence of tokens. Depending on the order of the entities, each instance can take one out of three gold standard relational labels, CONTAINS, CONTAINED-BY, and NONE.

The first line of Figure 1 is the token sequence for three gold standard entities, of which two are events, "surgery" and "scheduled", and one is a time expression, "March 11, 2014", whose time class is "date". One can form three candidate relations for these three entities.

3.2 Window-based processing

We aim to build a BERT-based model for both within- and cross-sentence relations. Figure 2 presents the distribution of the distance between the relation arguments in the THYME colon cancer training set expressed as tokens, e.g., 93.07% of the relation arguments are within 50 tokens; 95.14% are 60 tokens apart; 75% are within-sentence.

Thus, instead of looking for candidate pairs within a sentence, we look for pairs within a window of tokens of each other. We test window sizes of 50 or 60 tokens to balance coverage and good positive-to-negative ratio. By using a 60-token window and closure, we derive 413,327 NONE, 10,483

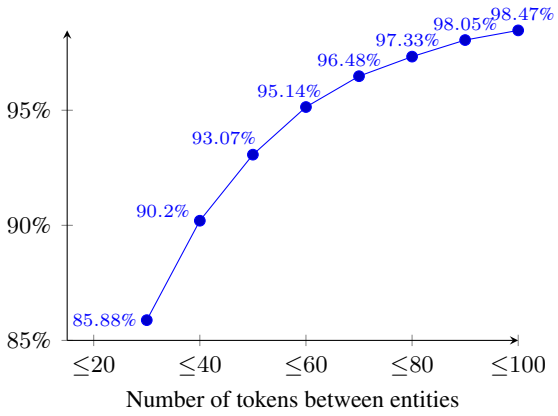

Figure 2: Relation coverage per token distance

CONTAINS, and 2,802 CONTAINS-BY instances from the THYME colon training set. Specifically, for every pair of entities[1] within a section (or if the document is not sectioned, every pair of entities within the document), we generate a relational candidate if the number of base tokens between the entities in the pair is less than the set window size.

XML-tags are often used to mark the position of the entities under consideration in a candidate pair (Dligach et al., 2017), and time expressions with their time class (Lin et al., 2017, 2018) for better generalizability. BERT uses the WordPiece tokenizer which breaks the XML-style tags (especially delimiters like angle brackets and slashes) into subtags. Therefore, we use non-XML tags to mark the positions of the entities and to encode time classes. Such tags should not be actual words and should not be broken into many tokens by WordPiece. Per the case in Figure 1, the event in an event-time relation pair is marked by **es** (event start) and **ee** (event end) and the time expression is represented by non-XML tags (**ts** for time start and **te** for time end) and its time class, for example **ts date te**. Event-event instances are marked with **eas** for event A start, **eae** for event A end, **ebs** for event B start, and **ebe** for event B end, for example . *a **eas** surgery **eae** is **ebs** scheduled **ebe** on march* 11.

3.3 BioBERT and BERT-MIMIC

A recent publication describes pre-training of BERT on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) (BioBERT; Lee et al., 2019).[2] We took this approach a step further and pre-trained BERT on clinical data from the MIMIC-III (Medical Information Mart for Intensive Care) dataset (Johnson et al., 2016). MIMIC-III contains 879 million words of patients' electronic medical records from Beth Israel Deaconess Medical Center's Intensive Care Unit. The resulting BERT-MIMIC model encapsulates clinical-domain-specific representations.

3.4 Augmenting with "silver" instances

Lin et al. (2018) describe a self-training routing in which they applied a model trained on the labeled THYME data to generate predictions on a set of unlabeled colon cancer data to create "silver" annotations. They demonstrated that adding high confidence positive "silver" relations to the gold training set improves the neural model performance. We apply this technique to our BERT-based models. The differences are 1) our unlabeled colon cancer instances are generated through the window-based mechanism, while their unlabeled instances were sentence-based; 2) we use a fine-tuned BERT model for generating "silver" instances.

3.5 Settings

We use a single NVIDIA GTX Titan Xp GPU to pre-train BERT on MIMIC-III, and fine-tune BERT, BioBERT, and BERT-MIMIC for our task. We use $\text{BERT}_{\text{base}}$, as the memory requirement of $\text{BERT}_{\text{large}}$ is too demanding. For fine-tuning, the batch size is selected from (16,32) and the learning rate is selected from (1e-5, 2e-5, 3e-5, 5e-5), using the THYME colon cancer development set. The fine-tuning is done with the Tensorflow-based BERT API, with the hidden state of the "[CLS]" token as the input to the classification layer. Rather than pre-training from scratch, which requires significant computational resources and would remove potentially useful information from the model, we initialize the pre-training on MIMIC data from BERT's final check point, with 10,000 training steps, standard warm up, and takes three hours to finish.

4 Results

All models are evaluated by the standard Clinical TempEval evaluation script so that their performance can be directly compared to published results. Table 1 shows performance on the Clinical TempEval colon cancer test set for the previous best systems, Lin et al. (2018) and Galvan et al. (2018), and window-based universal models.

[1] we use the term "entity" to refer to events and time expressions

[2] BioBERT model available at https://github.com/naver/biobert-pretrained

Model	P	R	F1
Lin et al. (2018)	0.692	0.576	0.629
Galvan et al. (2018)	**0.983**	0.462	0.629
1. bi-LSTM	0.712	0.490	0.581
2. BERT	0.699	0.625	0.660
3. BERT-T	0.735	0.613	0.669
4. BERT-TS	0.670	**0.697**	0.683
5. BioBERT(pmc)-TS	0.674	0.695	**0.684**
6. BERT-MIMIC-TS	0.673	0.686	0.679

Table 1: Model performance of *CONTAINS* relation on colon cancer test set. T: using non-XML tags; S: adding high confidence positive silver instances.

Model	P	R	F1
Lin et al. (2018)	**0.514**	0.585	0.547
BERT-TS	0.456	0.704	0.553
BioBERT(pmc)-TS	0.473	0.700	**0.565**
BERT-MIMIC-TS	0.457	**0.715**	0.558

Table 2: Model performance of *CONTAINS* relation on brain cancer test set.

We feed the window-based instances with XML-tagged entities to the bidirectional LSTM model without self-training (Lin et al., 2018) (Table 1(1)) as a comparison. Window-based instances with XML-tagged entities (Table 1(2)) and with non-XML tagged entities (Table 1(3)) are fed to BERT to show the difference from tagging. Then, high-confidence positive "silver" instances are added to the training set, fine-tuning is performed for BERT (Table 1(4)), BioBERT(pmc) (Table 1(5)) which showed better results than BioBERT trained on PubMed and PMC+PubMed, and BERT-MIMIC (Table 1(6)) respectively.

To evaluate the generalizability of the models, the best performing models trained on the colon cancer data – BERT (Table 1(4)), Bio-BERT(pmc) (Table 1(5)), and BERT-MIMIC (Table 1(6)) – are directly tested on the Clinical TempEval THYME brain cancer test set. Previous best cross-domain result is reported by Lin et al. (2018) in Table 2.

Thus, we establish a new state-of-the-art result for the task – 0.684F for within-domain (0.055 point improvement) and 0.565F for cross-domain (0.018 point improvement).

5 Discussion

The window-based BERT-fine-tuned model, even with the XML-tags (Table 1(2)), works for both within- and cross-sentence relations. Its performance (0.660F) is better than enhanced within-sentence models (Lin et al., 2018; Galvan et al., 2018) (0.629F), and the combination of two separate within- and cross-sentence models (Tourille et al., 2017) (0.613F). The improvement comes from 1) the window-based processing mechanism that bypasses the errors generated by a sentence boundary detector (for example, the sentence splitter creates two sentences for Figure 3(1) by incorrectly disambiguating the period after Mr); 2) the superb long-distance reasoning ability of BERT (Figure 3(2) shows relations we now can pick up from a three-sentence span). As a comparison, the same window-based approach does not work well with bidirectional LSTM model (Table 1(1)). One reason could be that because the bi-LSTM model is not pre-trained on a large corpus, it is likely affected by the limited number of gold annotations especially for large window sizes (like 50 or 60 tokens) which leads to skewing the positive/negative instance ratio further towards the negative labels, thus making fewer positive predictions (0.490 recall). Another explanation could be the different ways the bi-LSTM and BERT implement bidirectionality; each pass of the bi-LSTM is biased towards its nearby information thus favoring short-distance relations within a sentence.

The THYME corpus distribution does not provide gold sentence annotations. The BERT results we present in Table 1 are derived using a 60-token window. This window size produced superior results compared to a 50-token window (0.660F and 0.651F respectively).

Non-XML tags work better with BERT as they

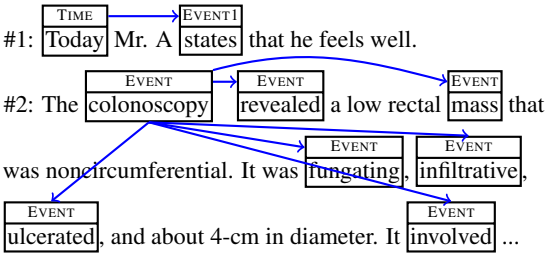

Figure 3: Relations picked up by the universal model.

Category	P	R	F1
within-sentence	0.621	0.712	0.663
cross-sentence	0.359	0.310	0.333

Table 3: Within- vs. cross-sentence results on colon cancer development set.

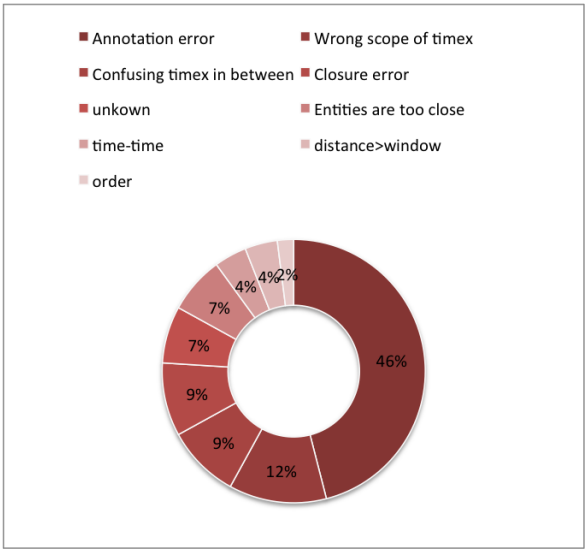

Figure 4: Distribution of 100 Errors

are not split into sub-tags but better preserved (Table 1(2)) vs. (Table 1(3)). We experimented with adding entity tags into BERT's vocabulary, instead of relying on strings (i.e. "es", "ee") that could possibly be confused with real tokens, but did not observe improved performance. We hypothesize that the BERT model needs to be re-trained with the added tags to contextualize their representations. Currently, we are limited by our computational resources to undertake such an endeavor.

Adding high quality silver instances is helpful as they alleviate the skewed positive to negative instance ratio, (Table 1(3)) vs. (Table 1(4)).

BERT-TS and its domain-specific versions (BioBERT(pmc)-TS, BERT-MIMIC-TS) work on par with each other (Table 1(4-5)) for in-domain tasks, and BioBERT(pms)-TS performs better when it is tested for generalizability on the the brain cancer Clinical TempEval test set (Table 2). The clinical-domain specific representation BERT-MIMIC-TS shows slight cross-domain advantage (0.558F) over BERT-TS (0.553F).

We performed error analysis on the output of the best performing model – BioBERT(pmc)-TS – on the THYME colon cancer development set. Applying this model results in 7.0k within-sentence CONTAINS predictions (4.3k correct) and 1.6k cross-sentence predictions (0.6k correct). Table 3 shows the within- and cross-sentence results of the best model on the colon cancer development set. However, these results should not be taken literally but as only an overall trend because closure over the entire set of relations needs to be factored, making it hard to isolate the performance of specific subtypes. For that reason, we did not subtype the results into event-event and event-time instances.

We sampled 100 errors evenly distributed over four categories: within-sentence false positives (FP), within-sentence false negatives (FN), cross-sentence FPs, and cross-sentence FNs. The sources of errors are summarized in fig. 4. 1) "Annotation error" (46%) – errors in the gold annotations; 2) "Wrong scope of timex" (12%) – the main reason for FP predictions, especially for cross-sentence ones (10%). The system fails to identify the subtle change of the timex scope and incorrectly links an event to it; 3) "Confusing timex in between" (9%) – there is another time expression occurring between the two arguments, thus the system incorrectly infers the scope of the time expression; 4) "Closure error" (9%); 5) "Unknown" (7%) – errors for which we could not provide a plausable explanation ; 6) "Entities are too close" (7%) – the two entities in question are too close to each other, thus limiting the context for correct reasoning. Prior knowledge would be helpful for these short-distance relations; 7) "Time-time" (4%) – the system generates time-time relations which are oftentimes FPs because gold time-time annotations are scarce ; 8) "Distance > window" (4%) – the distance between the two entities in question is bigger than the window size, resulting in cross-sentence FNs ; 9) "Order" (2%) – the system incorrectly extracts the order of the relation arguments, e.g. predicts CONTAINS instead of CONTAINS-BY .

One path for future research is pre-training BERT on a much larger clinical corpus (for which large scale computational resources are needed). The PMC set may not be clinical enough and the size of MIMIC corpus (0.9B) is too small compared to the other corpora (PubMed 4.5B, PMC 13.5B) to provide sufficient representations.

Acknowledgments

The study was funded by R01LM10090, R01GM114355 and U24CA184407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Sean Finan for his technical support and the anonymous reviewers for their valuable suggestions and criticism. The Titan Xp GPU used for this research was donated by the NVIDIA Corporation.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 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.

Steven Bethard, Guergana Savova, Martha Palmer, James Pustejovsky, and Marc Verhagen. 2017. Semeval-2017 task 12: Clinical tempeval. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 563–570.

Amar K Das and Mark A Musen. 1995. A comparison of the temporal expressiveness of three database query methods. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 331. American Medical Informatics Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. *EACL 2017*, page 746.

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.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3.

Michael G Kahn, Larry M Fagan, and Samson Tu. 1990. Extensions to the time-oriented database model to support temporal reasoning in medical expert systems. *Methods of information in medicine*, 30(1):4–14.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

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, Elizabeth W Karlson, Dmitriy Dligach, Monica P Ramirez, Timothy A Miller, Huan Mo, Natalie S Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, and Guergana K Savova. 2014. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association*.

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.

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

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

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 M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf*.

Reinhold Schmidt, Stefan Ropele, Christian Enzinger, Katja Petrovic, Stephen Smith, Helena Schmidt, Paul M Matthews, and Franz Fazekas. 2005. White matter lesion progression, brain atrophy, and cognitive decline: the austrian stroke prevention study. *Annals of neurology*, 58(4):610–616.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 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, Aurelie Neveol, 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)*, volume 2, pages 224–230.

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. *arXiv preprint arXiv:1902.01030*.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical dataa review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202.