End-to-End Temporal Relation Extraction in the Clinical Domain

José Javier Saiz[1,*], Begoña Altuna[2]

[1]University of the Basque Country (UPV/EHU), Arriola Pasealekua 2, Donostia, Gipuzkoa, 20018, Spain
[2]HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU

Abstract

Temporal relation extraction is an important task in the clinical domain, as it allows a better understanding of the temporal context of clinical events. In this paper, we present an end-to-end temporal relation extraction system for the clinical domain, using the i2b2 2012 Temporal Relation challenge as a benchmark. In our proposal, we fine-tune REBEL—a sequence-to-sequence model for general relation extraction—with temporal annotations and discharge summaries. Our proposal is then able to simultaneously extract relevant clinical entities, time expressions and the temporal relations between them. Our results demonstrate the effectiveness of this approach, achieving reasonable performance on the End-To-End track of the i2b2 2012 Challenge.

Keywords

Temporal Information Extraction, End-To-End Relation Extraction, Electronic Health Records

1. Introduction

Patients' medical information is stored in Electronic Health Records (EHRs), which contain structured data (e.g. demographics, vital signs and test results) and free text, such as reports and discharge summaries. The latter is the most informative, but the free text format makes clinical narratives prone to information overload, redundancy and poor access to information. This leads to inefficiencies [1], and can ultimately impact patient care negatively.

Extracting structured information from free-text clinical narratives can improve information accessibility and enhance patient care by facilitating clinical workflow. For example, Temporal Relation Extraction (TRE), which involves identifying events and time anchors according to their temporal features and then classifying the relations between these entities, can be applied to clinical timeline summarisation and ICD-10 code ordering, thus aiding medical research and patient care.

However, clinical TRE presents several challenges. Clinical narratives often have a high density of technical information and a concise writing style, which can make language modelling

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'23 Workshop, Dublin (Republic of Ireland), 2-April-2023

*Corresponding author.

josejavier.saiz.anton@gmail.com (J. J. Saiz); begona.altuna@ehu.eus (B. Altuna)
https://github.com/jsaizant (J. J. Saiz)
0000-0002-6883-3549 (J. J. Saiz); 0000-0002-4027-2014 (B. Altuna)

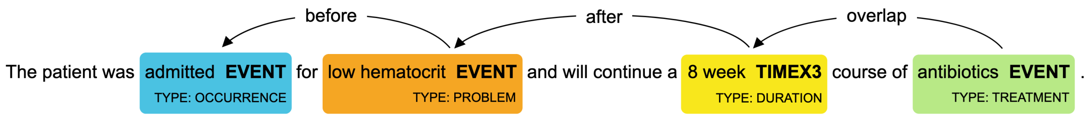


Figure 1: Sample sentence with entity and temporal relation annotations.

challenging [2]. In addition, the clinical lexicon and syntax can vary significantly across regions, institutions and medical specialties, making it difficult to develop universal approaches.

Our TRE system is designed to extract entities and temporal relations from clinical narratives (as shown in Figure 1). It is an end-to-end approach because it performs all TRE tasks simultaneously, namely temporal expression identification and temporal relation classification, as a sequence-to-sequence problem. To achieve this, we fine-tune REBEL (Relation Extraction By End-to-end Language generation) [3], a pre-trained model based on the BART (Bidirectional Auto-Regressive Transformer) architecture [4], which extracts triplet sequences from general domain text. We use the i2b2 Temporal Relation corpus, which contains clinical narratives annotated with temporal information, to train and evaluate our system. Fine-tuning helps to adapt the model to the specifics of clinical text and the task of extracting clinical temporal relations with a small amount of annotated data and training time.

2. Related Work

TRE approaches have evolved from rule-based systems to traditional machine learning systems with specialised classifiers and heuristics, and then to deep learning systems [5]. In the general domain, recent approaches incorporate advanced DNN-based models capable of learning high-level representations for TRE. These methods can include advanced neural language models such as BERT [6, 7], as well as graph-based architectures that capture the global structure of temporal relations in a text and are able to tackle document-level relation extraction [8, 9].

In the clinical domain, current state-of-the-art TRE systems are based on pre-trained BERT models or variants of this architecture [10]. However, recent approaches often lack flexibility and ease of use because they typically focus on a specific task, such as relation classification [11, 12, 13], and are limited to building temporal relations from gold standard entities. Furthermore, some only address a limited subset of relations [14, 15], such as explicit intra-sentence temporal relations, as seen in systems evaluated on the Direct Temporal Relations corpus from Lee et al. [16]. There is a need for further development of end-to-end strategies that can handle all types of temporal relations with larger dependencies, including cross-sentence and implicit temporal relations, which this work addresses.

3. Resources

In this section, we describe the dataset and approach chosen for the development of our clinical TRE system. We provide details on the dataset features, the representation of the input and output data, the model architecture and the limitations of the approach.

3.1. Dataset

The dataset used in this work was developed by the Informatics for Integrating Biology and the Bedside (i2b2) project for shared clinical NLP tasks and consists of 310 discharge summaries annotated with temporal information. Discharge summaries are divided into two main sections: clinical history (recent clinical history up to admission) and hospital course (hospital course and treatment plan after discharge). Both sections include annotations for temporal information based on the ISO-TimeML standard [17], including clinical events, time expressions, and temporal relations. On average, a discharge summary contains 86.6 events, 12.4 time expressions, and 176 temporal relations [18]. Table 1 provides statistics on the raw number of entities and types in the whole corpus.

Table 1
Quantitative description of the i2b2 Temporal Relation corpus.

Entity type	Joint count	Train set	Test set
Clinical event (EVENT)	30.212	16.610	13.731
Time expression (TIMEX3)	4.210	2.390	1.820
Temporal relation (TLINK)	61.940	34.204	27.736

The labels used in the dataset have types and attributes to properly represent and classify text tokens, as defined and described by Sun et al. [19]:

- The EVENT tag represents relevant events or states in the patient's clinical timeline, such as "follow-up" or "admission". It includes a "type" attribute to specify the event type and "modality" and "polarity" attributes to indicate the certainty and valuation of the event.
- The TIMEX3 tag is used for temporal expressions, including dates, durations, times, and frequencies, such as "March 21, 2021" or "3 hours". TIMEX3 tags' additional attributes are "val", that holds the normalized value of the time expression and "mod", that holds the time modifier value.
- The TLINK tag encodes three types of relations: overlap, before and after. TLINKs relate each event to the document's admission or discharge date ("Section Time TLINK"). In addition, TLINKs connect EVENTS and TIMEX3s within the same sentence or across multiple sentences ("Non-Section Time TLINKs") .

3.2. Model

To develop our system, we fine-tuned REBEL (Relation Extraction By End-to-end Language generation) [3], which is a pre-trained model based on the BART architecture. REBEL frames relation extraction as a sequence-to-sequence task and is trained to produce a sequence of relational facts from the input text. To represent relational facts, REBEL groups entities and relations into triplet sequences and represents them in a linear text string using special marker tokens. For example, consider the following text:

> "After the accident, the patient was admitted for surgery and rehab."

Here, we find two relational facts: "(accident, before, surgery)" and "(accident, before, rehab)", which are represented as the following sequence of triplets:

"<triplet> accident <prob> surgery <tret> before <prob> rehab <tret> before"

Entity marker tokens, enclosed in angle brackets, indicate the order of the entities in a relation: `<triplet>` indicates the start of a relation, while the subsequent tokens indicate whether the entity they accompany is a head or tail within the relation. We elaborate on this in section 3.3.

The triplet sequences, along with the corresponding context, are used as labels during training. During inference, the objective of the model is to extract a sequence of triplets representing the semantic relations contained in an input text.

Formally, there is a text x and a sequence of relations $y = (y_1, ..., y_n)$ with n being the length of y, that is, the proposed number of relations found in the input text x. The model must estimate the value of y, i.e. yield $\hat{y}$, that maximises the conditional probability $p(y|x)$. This probability can be decomposed as the joint probability of the sequence of relations involved, that is, as the product of the probabilities of generating the relation y_i conditioned on the text x and also on the previous relations found $y_{<i}$, as shown in expression (1).

$$\hat{y} = \arg\max_{y} p(y \mid x) = \arg\max_{y} p(y_1, ..., y_n \mid x) = \arg\max_{y} \prod_{i=1}^{n} p(y_i \mid y_{<i}, x) \tag{1}$$

By fine-tuning REBEL, our system would benefit from several advantages of the core architecture that are most appropriate for clinical TRE. For one, REBEL can handle longer context sequences (up to 1024 tokens) and extract relations that span multiple sentences. This is important for capturing document time and cross-sentence relations, which are common in clinical TRE corpora. Furthermore, REBEL does not require any entity annotation or pre-processing, unlike other systems that may rely on entity recognition or linking modules. Finally, REBEL can be easily adapted to specific domains and annotation schemes with little resources and time, as it is pre-trained on a large number of relations. This is also helpful because there are different annotation schemes for temporal relation extraction, and developing such specific systems from scratch would be expensive.

3.3. Data representation

REBEL is pre-trained on a distantly supervised dataset of 220 relation types generated by linking Wikidata entities and English Wikipedia abstracts. In order to fine-tune REBEL for the task of temporal relation extraction, the small size of the training dataset posed the challenge of accurately learning new entities and relations. Therefore, and given that REBEL was not trained with temporal relations, we adapted the annotations in the i2b2 2012 corpus by using triplet sequences with textual representations that the model learned during its pre-training phase.

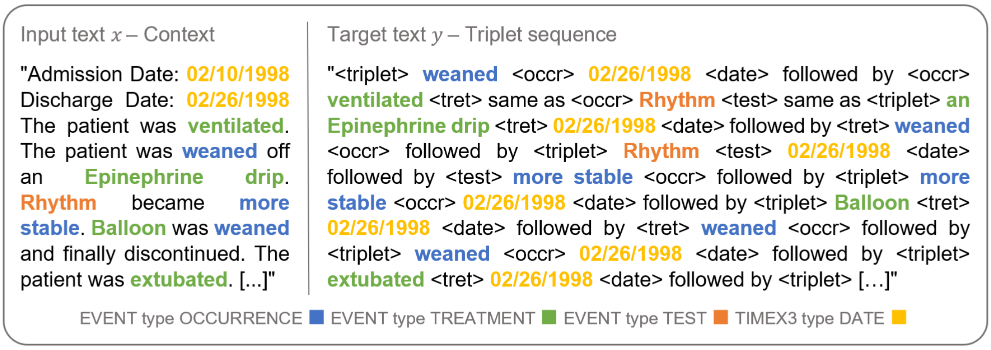
The relation types in the i2b2 2012 corpus were modified to match the textual form of the pre-training relations. The most appropriate textual representations were selected on the basis of semantic similarity and conciseness. For example, the relation type "after" was represented as "follows" and "before" as "followed by" in the triplet sequences. In the case of the "overlap" relation, we experimented with different representations such as "said to be the same as" or "partially coincident with". However, larger representations were found to propagate errors more easily, so we finally chose "same as".

We also modified the entity marker tokens. The original REBEL model used only three tokens to structure the triplet sequence: `<triplet>`, `<subj>` and `<obj>`. We added new tokens that indicated the entity types as well. That is, for each of the 10 entity types in the i2b2 2012 corpus, we created a new entity marker token with a four-character abbreviation. For example, `<tret>` for the EVENT type "treatment" and `<freq>` for the TIMEX3 type "frequency". We added these tokens to the embedding dimension to ensure that they were processed accurately.

4. Experiments

The following section describes the fine-tuning process and evaluation of our system. In order to be comparable with other systems, we followed the rules of the i2b2 2012 Challenge and used the split provided, which consists of a train set of 190 documents and a test set of 120 documents from the i2b2 Temporal Relation corpus. Our modeling setup consisted of the following steps:

1. We segmented the 190 documents into texts of 512 characters each. This segmentation method accounted for sentences with different lengths and facilitated a more consistent division. Moreover, this also enabled us to utilize the maximum amount of text that could fit within the memory constraints of our model. We also made sure that each text contained the admission and discharge dates at the beginning, so that we could also extract the Section Time TLINKs. This resulted in 1.139 training instances, each comprising an input-output text pair, as illustrated by figure 2.
2. We fine-tuned the pre-trained REBEL model by iteratively feeding the train instances.



Input text x – Context

"Admission Date: 02/10/1998 Discharge Date: 02/26/1998 The patient was ventilated. The patient was weaned off an Epinephrine drip. Rhythm became more stable. Balloon was weaned and finally discontinued. The patient was extubated. [...]"

Target text y – Triplet sequence

"<triplet> weaned <occr> 02/26/1998 <date> followed by <occr> ventilated <tret> same as <occr> Rhythm <test> same as <triplet> an Epinephrine drip <tret> 02/26/1998 <date> followed by <tret> weaned <occr> followed by <triplet> Rhythm <test> 02/26/1998 <date> followed by <test> more stable <occr> followed by <triplet> more stable <occr> 02/26/1998 <date> followed by <triplet> Balloon <tret> 02/26/1998 <date> followed by <tret> weaned <occr> followed by <triplet> weaned <occr> 02/26/1998 <date> followed by <triplet> extubated <tret> 02/26/1998 <date> followed by <triplet> [...]"

EVENT type OCCURRENCE 🟦 EVENT type TREATMENT 🟩 EVENT type TEST 🟧 TIMEX3 type DATE 🟨

Figure 2: Sample training instance consisting of a context and a triplet sequence containing the entities and relations in the context. Given input text x, the model is trained to minimize the difference between the predicted output and the target text y at each time step.

Our system was obtained at the last checkpoint after fine-tuning the base REBEL model for 10 epochs, and following the training parameters shown in Table 2. Then, the system was evaluated with the End-To-End track from the i2b2 2012 Challenge [18]. The evaluation setup consisted in the following process:

1. We divided the 120 documents in the test set into 512-character texts, each beginning with the admission and discharge dates.
2. We fed each text into the fine-tuned model, which conditionally generated text sequences of triplets.
3. We decoded the triplet sequences and compiled them into XML format, which were then evaluated using the i2b2 2012 Temporal Evaluation Scripts. No additional post-processing was performed.

Table 2
Training and inference parameters used to fine-tune and generate predictions with our system.

Training		Inference	
Parameter	Value	Parameter	Value
Optimization	AdamW	Strategy	Beam search
Embedding size	1024	Maximum length	1024
Learning Rate	0.00005	Length penalty	0
Epochs	10	Beam number	3
Batch size	1	Returned sequences	1

The BART architecture employs conditional text generation to generate predictions, which can be influenced by modifying various decoding parameters. We chose the Beam Search strategy among multiple decoding strategies available, as it produces more consistent results by exploring and scoring multiple output sequences in parallel [20]. Since the order of the entities and marker tokens affects the accuracy of the predictions, consistency of the predicted tokens is crucial to prevent errors from propagating to the subsequent sequence of text. The Beam Search strategy considers multiple output sequences simultaneously and generates subsequent tokens based on the top-k sequences with the highest scores, ensuring that the extracted relations are consistent with the input sequence. The parameters used to generate the predictions for evaluation are shown in Table 2. For reproducibility purposes, the code to replicate the dataset and modelling setup is available on our GitHub page under the CC BY-SA-NC 4.0 licence [1]. For confidentiality reasons, data and evaluation scripts are only available on request from the n2c2 organisation.

5. Results

To be consistent with previous benchmarks, we adopt the TempEval3 evaluation metrics used in the original i2b2 2012 challenge [18], which calculates the Precision, Recall and Micro-average F1 scores. Here, the evaluation metrics differ from the standard F1 used for standard multi-class settings in that Precision is computed by verifying each predicted relation against the transitive closure of the gold standard, and Recall is computed by verifying each gold standard relation against the transitive closure of the predictions. Table 3 shows our system's results relative to those of the best performing systems in the i2b2 2012 challenge.

[1] https://github.com/jsaizant/ETEREX-REBEL

Table 3
Performance of the top 3 scoring systems in the End-To-End track of the i2b2 2012 Challenge against our proposal (in italics). TLINK F-score is the primary measure.

System	TLINK			EVENT		TIMEX3	
	F1	P	R	Span F1	Type P	Span F1	Type P
Tang et al. [21]	**0.63**	**0.7**	0.57	0.9	0.84	0.87	0.85
Xu et al. [22]	0.59	0.59	**0.59**	**0.92**	**0.86**	**0.91**	**0.88**
Our proposal	*0.58*	*0.65*	*0.52*	*0.78*	*0.72*	*0.77*	*0.65*
Roberts et al. [23]	0.53	0.48	0.57	0.89	0.8	0.89	0.78

Our system delivered mixed results when tested on the i2b2 2012 dataset. In the EVENT and TIMEX3 extraction tasks, the system's accuracy was 0.78 and 0.77 respectively, which is below the best performing systems. However, in the TLINK extraction task, the system achieved an accuracy of 0.58, ranking third among the best performing systems. In terms of architecture, the other three systems use a pipeline approach consisting of multiple CRF and SVM classifiers and rule-based methods for time expression detection and/or normalisation [21, 22, 23]. It is also worth noting that while our experiments are limited to the i2b2 2012 corpus, they used additional corpora to build their event classifiers: Tang et al. [21] and Xu et al. [22] used the i2b2 2010 corpus and Roberts et al. [23] used additional text resources from PubMed, Wikipedia and other medical records. Comparatively, our training data is more restricted, which serves to demonstrate the adaptability of our approach.

Despite its poor performance on the EVENT and TIMEX3 tracks, the system's performance on the TLINK extraction is relatively stronger, suggesting that it excels at recognising temporal relations between entities rather than recognising the entities themselves. The results are in line with expectations for a system performing the event, time expression and temporal relation tasks simultaneously. The TRE task has traditionally consisted of a first subtask of entity extraction and a second subtask of relation extraction, and high performance in the first was crucial for good results in the second. Our system, fine-tuned directly for relation extraction, does not seem to be so dependent. However, improving the system's ability to recognise clinical and temporal entities could potentially improve its performance in the TLINK extraction task.

In accordance with the overall extraction results, the remarkably low results in the classification of event and time expression types do not seem to be crucial for the temporal relation extraction task. However, taking into account the semantic information encoded in the entity type could also help in the classification of temporal relations.

6. Discussion

Our system uses the REBEL framework for Temporal Relation Extraction, a high-level NLP task that requires contextual temporal information. Since REBEL uses the BART architecture, the bi-directional encoder and attention mechanism allows the model to focus on specific parts of the input text and weigh their importance in relation to the task at hand, giving the model a greater capacity to process longer texts. The system's capabilities have been further enhanced by a fine-tuning process using contextual texts and triplet sequences with lengths of up to 1024 embedded tokens. These longer sequences allow the temporal context of the information to be better captured.

The i2b2 2012 corpus contains the "BEFORE" type as the most frequent TLINK type. To assess the performance of our proposal, we perform an evaluation using a test set consisting of pairs of entities predicted by the model and the most frequent TLINK class, i.e. the "BEFORE" type. We refer to this evaluation as the baseline evaluation, and its purpose is to determine whether the performance of the model is influenced by the most frequent class, and whether this influence affects the performance on the other classes. The baseline score shows precision and recall values of 0.50 and 0.25 respectively, and its F-score is 0.33, which is 0.25 lower than the F-score of our model prediction. Although the proposed model performed better than the baseline evaluation, the difference in F-scores was not significant. This suggests that the model is indeed biased by the most frequent TLINK class and that there is still potential to improve the overall performance of the model. While the model successfully identified most of the entity pairs, the classification of the TLINK type was challenging. Indeed, most of the narratives tend to be written in the past tense, which favours the temporal order in a certain direction. This tendency creates a class imbalance in the temporal annotations, leading to a bias in the model's predictions, which is where further work is most needed and what we discuss next.

Despite the large number of temporal annotations in the training dataset, not all the possible relations are considered, making it difficult to train a system that produces consistent predictions. For example, consider the set of entities in Figure 3. There may be several ways to label the relations between them, such as "A" before "B" and "B" before "C", or "B" after "A" and "C" after "B". In both cases, the relation between "A" and "C" remains unlabelled because not all relations are explicitly annotated. This carries over to system inference, where we find that reciprocal relations are not identified at all, and transitive relations are usually overlooked.

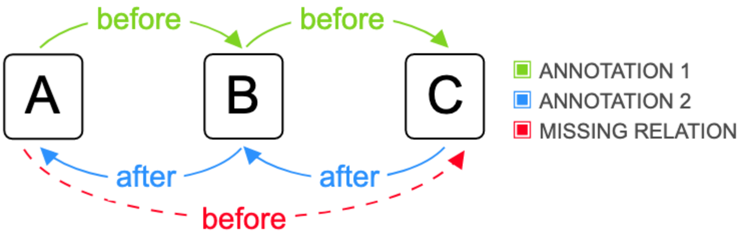


Figure 3: Under the i2b2 2012 annotation guidelines, reciprocal relations can be explicitly labelled in multiple ways, and transitive relations are often omitted.

To overcome the sparsity of TLINK annotations in the training set and to improve the performance of the system, we propose two solutions. First, we propose the use of transitive closure, which derives implicit relations from existing labelled relations, thereby increasing the number of annotations. Together with the integration of reciprocal relations into the training instances, this approach has been shown to mitigate the imbalance of TLINK types and improve system performance [24]. Secondly, we propose to use a training corpus with narrative container annotations [25], such as the E3C corpus [26], which prevents redundant relations, but also reduces the distance of temporal dependencies, narrowing the context needed for relation identification.

In section 5 we observed that the extraction of relations in our system is less dependent on the extraction of event and time expressions than traditional TRE systems. However, better entity identification could improve the overall performance. For sequence-to-sequence architectures such as our system, it is crucial to find the most effective textual representation of information. As explained in section 3.3, we adapted the annotations of the i2b2 2012 corpus by matching TLINK types to similar pre-training relations and creating abbreviated entity markers, which proved to be a successful adaptation. To further improve entity extraction, a potential solution is to assign token embeddings from the pre-trained model to the new entity markers. This will improve entity extraction by using existing weights and biases for entity markers such as `<tret>` and `<freq>` rather than training new weights from scratch. Indeed, further exploration of different token representation techniques is needed to understand their impact on model performance.

7. Conclusion

This article presents our end-to-end system for clinical Temporal Relation Extraction, developed by fine-tuning REBEL with temporal annotations and discharge summaries from the i2b2 Temporal Relation corpus. Our system uses a sequence-to-sequence approach that annotates raw clinical narratives and extracts relations between entities in a single step, rather than relying on multiple mechanisms and highly engineered linguistic features. This makes our system more (re)usable and less dependent on specific texts or tasks. We have evaluated our system in the End-To-End track of the i2b2 2012 Challenge and achieved reasonable results, showing that our approach can handle complex clinical domains with limited resources and time. We plan to explore ways to improve our system in future work, such as using pre-trained token embeddings for the entity markers in the triplet sequences and training the system on a corpus of narrative containers.

Acknowledgments

This work has been partially funded by the Basque Government postdoctoral grant POS 2022 2 0024. We would also like to thank Arantza Casillas and Alicia Pérez for their advice and help.

References

[1] A. Mathioudakis, I. Rousalova, A. A. Gagnat, N. Saad, G. Hardavella, How to keep good clinical records, Breathe 12 (2016). doi:10.1183/20734735.018016.

[2] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li, et al., Pre-trained Language Models in Biomedical Domain: A Survey from Multiscale Perspective, arXiv e-prints (2021).

[3] P. L. H. Cabot, R. Navigli, REBEL: Relation Extraction by End-to-end Language generation, Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021 (2021). doi:10.18653/v1/2021.findings-emnlp.204.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehensio, Computer Methods and Programs in Biomedicine (2020). doi:10.18653/v1/2020.acl-main.703.

[5] Y. B. Gumiel, L. E. S. E. Oliveira, V. Claveau, N. Grabar, E. C. Paraiso, C. Moro, D. R. Carvalho, Temporal Relation Extraction in Clinical Texts, ACM Computing Surveys 54 (2022). doi:10.1145/3462475.

[6] S. Zhang, Q. Ning, L. Huang, Extracting Temporal Event Relation with Syntax-guided Graph Transformer, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 379–390. URL: https://aclanthology.org/2022.findings-naacl.29. doi:10.18653/v1/2022.findings-naacl.29.

[7] X. Zhao, S. Lin, G. Durrett, Effective Distant Supervision for Temporal Relation Extraction, CoRR abs/2010.12755 (2020). URL: https://arxiv.org/abs/2010.12755. arXiv:2010.12755.

[8] B. Su, S. Hsu, K. Lai, A. Gupta, Temporal Relation Extraction with a Graph-Based Deep Biaffine Attention Model, CoRR abs/2201.06125 (2022). URL: https://arxiv.org/abs/2201.06125. arXiv:2201.06125.

[9] X. Xu, T. Gao, Y. Wang, X. Xuan, Event temporal relation extraction with attention mechanism and graph neural network, Tsinghua Science and Technology 27 (2022) 79–90. doi:10.26599/TST.2020.9010063.

[10] A. L. Olex, B. T. McInnes, Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be, Journal of Biomedical Informatics 118 (2021). doi:10.1016/j.jbi.2021.103784.

[11] C. Lin, T. Miller, D. Dligach, F. Sadeque, S. Bethard, G. Savova, A bert-based one-pass multi-task model for clinical temporal relation extraction, Association for Computational Linguistics (2020). doi:10.18653/v1/2020.bionlp-1.7.

[12] Y. Zhou, Y. Yan, R. Han, J. H. Caufield, K.-W. Chang, Y. Sun, P. Ping, W. Wang, Clinical Temporal Relation Extraction with Probabilistic Soft Logic Regularization and Global Inference, 2020. URL: https://arxiv.org/abs/2012.08790. doi:10.48550/ARXIV.2012.08790.

[13] H. U. Haq, V. Kocaman, D. Talby, Deeper Clinical Document Understanding Using Relation Extraction, 2021. arXiv:arXiv:2112.13259.

[14] H. Guan, J. Li, H. Xu, M. Devarakonda, Robustly Pre-trained Neural Model for Direct Temporal Relation Extraction, 2020. arXiv:arXiv:2004.06216.

[15] G. Alfattni, N. Peek, G. Nenadic, Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries, Journal of Biomedical Informatics 123 (2021) 103915. URL: https://www.sciencedirect.com/science/article/pii/S1532046421002446. doi:https://doi.org/10.1016/j.jbi.2021.103915.

[16] H.-J. Lee, Y. Zhang, M. Jiang, J. Xu, C. Tao, H. Xu, Identifying direct temporal relations between time and events from clinical notes, BMC Medical Informatics and Decision Making 18 (2018). URL: https://doi.org/10.1186/s12911-018-0627-5. doi:10.1186/s12911-018-0627-5.

[17] J. Pustejovsky, K. Lee, H. Bunt, L. Romary, ISO-TimeML: An international standard for semantic annotation, Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010 (2010).

[18] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012

i2b2 Challenge, Journal of the American Medical Informatics Association 20 (2013). doi:10.1136/amiajnl-2013-001628.

[19] W. Sun, A. Rumshisky, O. Uzuner, Annotating temporal information in clinical narratives, Journal of Biomedical Informatics 46 (2013). doi:10.1016/j.jbi.2013.07.004.

[20] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, J. Weston, Neural Text Generation with Unlikelihood Training, CoRR abs/1908.04319 (2019). URL: http://arxiv.org/abs/1908.04319. arXiv:1908.04319.

[21] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, H. Xu, A hybrid system for temporal information extraction from clinical text, Journal of the American Medical Informatics Association 20 (2013) 828–835. URL: https://doi.org/10.1136/amiajnl-2013-001635. doi:10.1136/amiajnl-2013-001635.

[22] Y. Xu, Y. Wang, T. Liu, J. Tsujii, E. I.-C. Chang, An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge, Journal of the American Medical Informatics Association 20 (2013) 849–858. URL: https://doi.org/10.1136/amiajnl-2012-001607. doi:10.1136/amiajnl-2012-001607.

[23] K. Roberts, B. Rink, S. M. Harabagiu, A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text, Journal of the American Medical Informatics Association 20 (2013) 867–875. URL: https://doi.org/10.1136/amiajnl-2013-001619. doi:10.1136/amiajnl-2013-001619.

[24] G. Alfattni, N. Peek, G. Nenadic, Extraction of temporal relations from clinical free text: A systematic review of current approaches, Journal of Biomedical Informatics 108 (2020). doi:10.1016/j.jbi.2020.103488.

[25] J. Pustejovsky, A. Stubbs, Increasing Informativeness in Temporal Annotation, in: Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 152–160. URL: https://aclanthology.org/W11-0419.

[26] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C project: Collection and annotation of a multilingual corpus of clinical cases, CEUR Workshop Proceedings 2769 (2020). doi:10.4000/books.aaccademia.8663.