Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 349-362 recibido 05-12-2024 revisado 22-01-2025 aceptado 03-02-2025

Do Entailment Models know about Reasoning Temporal Ordering on Clinical Texts?

¿Saben razonar los modelos de vinculación sobre el orden temporal en textos clínicos?

Edgar Andres Santamaria[1], **Oier López de Lacalle**[2]
Aitziber Atutxa[1], **Koldo Gojenola**[1]

[1]Bilbao School of Engineering, [2]San Sebastian Faculty of Computer Science,
HiTZ Basque Center for Language Technology
University of the Basque Country (UPV/EHU), Spain
{edgar.andres, oier.lopezdelacalle, aitziber.atucha, koldo.gojenola}@ehu.eus

Abstract: This paper investigates the use of entailment-based methods for Clinical Temporal Relation Extraction (CTRE), addressing challenges such as data scarcity, label imbalance, and domain-specific complexity. By reframing the task as a Natural Language Inference (NLI) problem, the approach reduces annotation requirements and improves generalization across datasets. Experiments with the THYME and E3C corpora show that NLI-based models outperform traditional classifiers in low-resource settings, demonstrating strong transferability and resilience to class imbalance, making them an effective solution for CTRE in clinical narratives.
Keywords: Temporal Relation extraction, Natural Language Inference, Medical domain, transfer learning

Resumen: Este artículo investiga el uso de métodos basados en Inferencia del Lenguaje Natural (NLI) para la extracción de relaciones temporales en textos clínicos (CTRE), abordando desafíos como la escasez de datos, el desequilibrio de las etiquetas y la complejidad específica del dominio. Al reformular la tarea como un problema de NLI, el enfoque reduce los requisitos de anotación y mejora la generalización entre conjuntos de datos. Los experimentos con los corpus THYME y E3C muestran que los modelos basados en NLI superan a los clasificadores tradicionales en entornos de bajos recursos, lo que demuestra una fuerte transferibilidad y robustez frente al desequilibrio de clases, lo que los convierte en una solución eficaz para CTRE en narrativas clínicas.
Palabras clave: Extracción de relaciones temporales, inferencia del lenguaje natural, dominio médico, aprendizaje por transferencia

1 Introduction

In medical reasoning, the temporal evolution of the events related to a patient's situation is crucial for diagnosis, prognosis or even for making the right therapeutic decisions. Several studies have investigated the impact that temporal information might have on different systems and medical contexts, such as inclusion criteria on clinical trials, chronic disease risk prediction or patient phenotyping, among others (Dalianis, 2018), (Bui, Aberle, and Kangarloo, 2007), (Hirsch et al., 2014), (Caron et al., 2017), (van der Linden, van Wijk, and Funk, 2021), (Wang et al., 2022). For example, in Intensive Care Units (ICU), evolution notes store the temporal progression of a medical condition of a patient indicating the amelioration or deterioration after the administration of a given treatment, or the application of a given procedure. Knowing temporal dependencies between related events is especially relevant in this case (Johnson et al., 2016).

Clinical Timeline Extraction is a high-level task that aims to build temporal representations of clinical texts. The task is usually decomposed in three lower-level tasks: 1) Identification of temporal expressions and clinical events, 2) extraction of temporal relations among clinical events and temporal expressions, and 3) clinical event timeline ordering. In this paper, we will focus on the second task.

ISSN 1135-5948 DOI 10.26342/2025-74-24

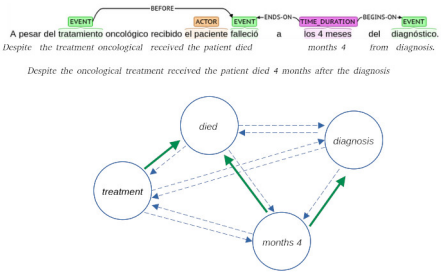

Figura 1: Dashed arrows represent missing-relations (i.e. non-anotated relations) either because they are implicit and derivable (e.g. diagnosis *before* died) or underspecified (Missing-Relations). Green arrows represent explicitly annotated relations.

Temporal reasoning is challenging in two respects. First, initially, each event or time expression could potentially be combined with respect to all the others, leading to a substantial increase in hypotheses, which scales quadratically with respect to the total number of entities (events and time expressions). However, as (Ning, Feng, and Roth, 2017) pointed out, the majority of conceivable event-time expression pairings lack any explicit connection in the manual annotation (see Figure 1). This has as a consequence a highly imbalanced number of actually related pairs and annotated as such compared to the missing ones.

The second challenge, as most recent reviews point out (Alfattni, Peek, and Nenadic, 2020),(Olex and McInnes, 2021), concerns specifically the medical domain. All studies on temporal relation extraction in clinical and medical narratives highlight that one of the major challenges is the lack of sufficiently large open-access datasets.

Data annotation requires expertise and time to annotate training data at large scale with sufficient consistency, which is very costly. In addition, annotated datasets show poor transfer properties between domains: Information Extraction (IE) annotations depend on the labeling schema used in each domain, and moving to new datasets requires new schemas, and the manual annotation of new data.

Entailment-based approaches have been widely used as an effective method to save annotation effort and perform transfer learning across multiple data sources and schemas (Sainz et al., 2021; Sainz et al., 2022; Baucells et al., 2023; Vashishtha et al., 2020). The approach consists of reformulating the classical Relation Extraction multiclass classification task (See Table 1) as a natural language inference (NLI) task and passing the input to an entailment model, whose output is then mapped to that of the target task.

Entailment-based models have primarily been applied to tasks requiring knowledge of general domains and factual relationships between entities (e.g., "Barack Obama was born in Hawaii"). However, their potential usefulness in specialized scenarios, such as those in the medical domain, has been underexplored. Similarly, the effectiveness of modeling temporal relations via entailment is still an open research question. Our research aims to investigate the feasibility of potential improvements of the temporal relation extraction using entailment models. We aim to show that the models are particularly effective in scenarios where there is little annotated data, where the annotated labels are extremely unbalanced and the domain is very specific (i.e. medical health records).

In this paper we present an entailment-based system for temporal relation extraction. We empirically show that (1) recasting the task as an entailment problem is an effective approach to model complex temporal relationships in the medical domain, as well as (2) there is a transfer for temporal knowledge between datasets. (3) We show that entailment-based models need less annotated data than supervised classifiers in scenarios of low data-regimes, and they are not that affected due to the extremely unbalanced distribution of the training examples. (4) We show that generic NLI models contain little temporal knowledge, but it can be learned from existing data sources.

2 Related Work

Relation Extraction (RE) aims to identify and extract complex relationships between entities referenced within textual content. Clinical Temporal Relation Extraction is a specific case of RE where the entities are clinical events and time expressions. The classical approach to RE conceives the task as a classification problem. Given a pair of entities $e1$ and $e2$ within the context X, the objective (see Equation 1) is to maximize the probability assigned to the correct class y (the actual relation). The set of classes includes all predefined relations, along with an additio-

Element type	Example	Relation type
EVENT-TIMEX	39.6 weeks of gestation	OVERLAP
EVENT-EVENT	history was negative for blood coagulopathies	CONTAINS
TIMEX3-TIMEX3	The patient was diagnosed with diabetes in 2018 and started insulin therapy two months later	BEFORE

Tabla 1: TLINKs among: temporal expressions (TIMEX3) and clinical events (EVENT).

nal OUTREL class, which accounts for both unannotated missing relations and true non-relations.

$$\hat{y} = \arg\max_{y \in C} P(y \mid e1, e2, X) \quad (1)$$

Since the irruption of the transformer architectures (Vaswani et al., 2017), approaches to solve both general domain and clinical Temporal Relation Extraction predominantly rely on pre-trained language models like BERT (Devlin et al., 2019) and special token embeddings to facilitate classification (Classical-BERT-based RE). Works on these lines include (Ning, Subramanian, and Roth, 2019; Lin et al., 2019; Zhou et al., 2021; Lin et al., 2021a; Lin et al., 2023; Knez and Žitnik, 2024). (Lin et al., 2019) introduce special tokens to mark events and employ a BERT network, using the [CLS] token embedding to classify temporal relations. (Zhou et al., 2021) improve performance through soft logic regularization, predicting relation probabilities while ensuring consistency with rule-based constraints. (Lin et al., 2021a) enhance extraction in the medical domain by utilizing entity-specific pre-training of BERT. Finally, (Knez and Žitnik, 2024), following the work done by (Lin et al., 2023) for medical entity factual relations, explore a bimodal architecture integrating information from text documents and knowledge graphs combining them to make a unified prediction.

Classical BERT-based RE systems require large amounts of labeled examples which are costly to annotate as addressed by (Sainz et al., 2021). This is especially relevant in the clinical domain because open-access datasets are scarce and their annotation requires specialized professionals. Additionally, classical BERT-based RE systems suffer enormously in datasets with imbalanced label distributions, since imbalance tends to drag the threshold towards the large class, that is, the OUTREL class (Wang et al., 2023). (Obamuyide and Vlachos, 2018; Sainz et al., 2021) demonstrate that reformulating RE as a Natural Language Inference (NLI) task helps reduce the annotation effort in low-data regimes where classical BERT-based RE techniques fail to perform accurately (Baldini Soares et al., 2019). Moreover, (Sainz et al., 2022; Baucells et al., 2023) demonstrated that reframing tasks as entailment problems reduces reliance on schemas, a primary barrier to transferring annotations across domains and datasets. Furthermore, adopting the NLI approach helps address the OUTREL imbalance issue.

Textual Entailment Textual Entailment focuses on reasoning about text relationships by determining the logical relationship between two pieces of text: a premise and a hypothesis. Originally introduced by (Dagan, Glickman, and Magnini, 2006), the task was later expanded upon by (Bowman et al., 2015), who termed it as Natural Language Inference (NLI). Given a textual premise and a hypothesis, the task is to decide whether the premise entails or contradicts (or is neutral to) the hypothesis. For RE, the premise corresponds to the context X, and the reformulation involves generating, by means of concrete verbalizations, a hypothesis for each possible relation $y \in C$ (see Figure 2).

Several studies have demonstrated that reformulating traditional classification tasks as Natural Language Inference (NLI) problems can effectively harness the reasoning capabilities of NLI models. This approach has been shown to improve performance and enhance generalization in the original classification tasks, particularly in scenarios with limited data availability (Uppal et al., 2020), (Sainz et al., 2021). As mentioned above, NLI reformulation involves leveraging high-level semantic reasoning in specific semantic tasks. For example, (Poliak et al., 2018) described efforts to recast 7 semantic phenomena, including RE, from a total of 13 datasets into NLI examples.

But to our knowledge temporal RE recast

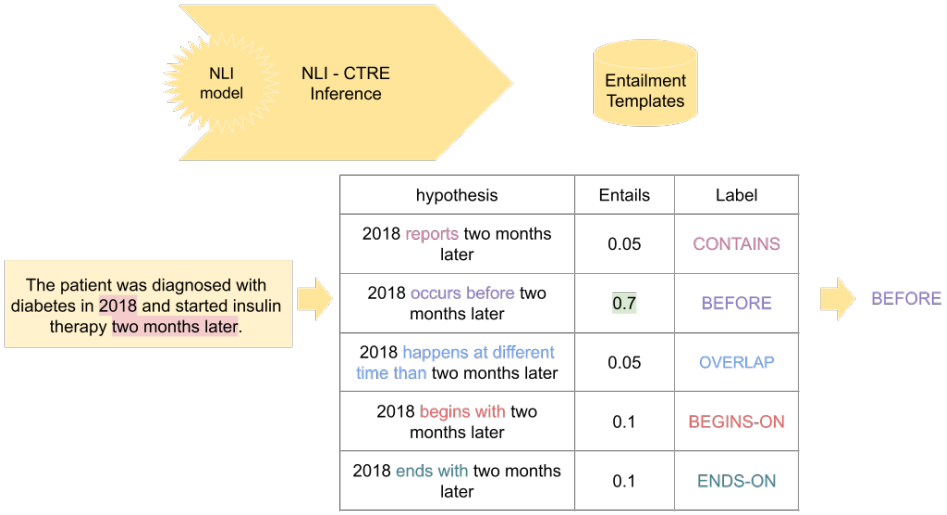

Figura 2: General overview of the inference using the NLI approach. The input text containing two entity mentions (on the left) is taken as the premise. The entailment templates are used to verbalize hypotheses using the entity mentions as placeholders (*2018* as $e1$ and *diagnosed* as $e2$ of the relation). The entailment probability of each premise-hypothesis pair is obtained from the NLI model, and the pair with the highest entailment probability is selected to infer the correct relation label. Note that each hypothesis is related to one label.

as NLI has only been explored by (Vashishtha et al., 2020). The authors reformulated multiple datasets annotated for event duration and event ordering into more than one million NLI examples and investigated how well models trained on generic NLI datasets capture these forms of temporal reasoning. Unlike our work, their temporal reasoning was limited to the `before` relation type.

3 Problem Formulation

Clinical Timeline Extraction involves constructing temporal representations of clinical texts, where TLINKs encode the temporal relations between the combination of two elements of TIMEX3 (temporal expressions) and EVENT (clinically relevant entities) as shown in Table 1. Most annotation guidelines are based on Allen's seven seminal temporal relations (Allen, 1983). In this paper, we focus specifically on the Clinical Temporal Relation Extraction (CTRE) subtask, centered on classifying each temporal relation (TLINK) within the label set C. The set C in most datasets comprises the following temporal relation labels: `BEFORE`, `BEGINS-ON`, `ENDS-ON`, `CONTAINS`, `OVERLAP`, and `OUTREL`.

The TLINK identification task is usually formulated as a classical relation extraction (RE) problem, specifically a pairwise classification task, where the objective is to determine the temporal relationship between every pair of EVENTS and TIMEX3s. This strategy has a quadratic cost since it generates TLINK candidates for all possible pairs of entities and produces a very unbalanced proportion of relation/`OUTREL` candidates as shown in Figure 1.

4 Entailment Based Approach

Our application of the entailment-based approach aligns with previous works (Sainz et al., 2022) and can be summarized into three steps illustrated in Figure 2. To reformulate Clinical Temporal Relation Extraction (CTRE) as an entailment task, each context X, which contains the entities $e1$ and $e2$, is treated as the premise. For each possible temporal relation, we generate a corresponding hypothesis.

This process involves crafting verbalizations that effectively represent the semantic meaning of the relations. For example, a verbalization for the `OVERLAP` relation might state that $e1$ occurs at the same time as $e2$, which would *entail* the existence of an `OVERLAP` relation between $e1$ and $e2$. Additionally, verbalizations that contradict each relation must also be generated. Neutral verbalizations are also employed, but for the sake of clarity, their generation will be explained later.

The task then involves determining the entailment probability for each plausible re-

Label	Entailment templates	Contradiction templates
BEFORE	{e1} occurs before {e2} {e1} happens before {e2} {e2} occurs after {e1} {e2} happens after {e1}	{e2} is prior to {e1} {e1} is subsequent to {e2}
CONTAINS	{e1} reports {e2} {e1} demonstrates {e2} {e1} shows {e2} {e1} is observed in {e2} {e1} objectifies {e2} {e1} confirms {e2}	{e2} consists of {e1} {e2} controls {e1} {e2} practices {e1}
OVERLAP	{e1} and {e2} occur at same time at some point in time {e1} and {e2} overlap at some point in time	{e1} happens at different time than {e2} {e1} shares no time with {e2}
BEGINS-ON	{e1} begins when {e2} begins {e1} begins with {e2} {e1} simultaneously begin {e2}	{e1} starts before {e2} {e2} starts before {e1} {e2} prior to {e1} {e1} Prior to {e2}
ENDS-ON	{e1} ends with {e2} {e1} ends when {e2} ends {e1} terminates with {e2}	{e2} terminates with {e1} {e2} ends with {e1}

Tabla 2: Verbalization templates for entailment and contradiction alignment for given {e1} (source) and {e2} (destination) entities (event and temporal expression mentions).

lation between $e1$ and $e2$ within the context X, relative to the contradiction and neutral probabilities independently for each relation. The final prediction is made by selecting the relation associated with the highest entailment probability. As described below, the entailment model can be further trained on temporal reasoning datasets by reformulating temporal RE as an entailment problem. Figure 2 shows the main workflow of the approach. First, given the input text that contains the two entities (i.e. TIMEX3 or EVENT mentions), all the **entailment verbalizations** are generated as the hypotheses for the entailment model. Note that the input context is considered as the premise for the model. Second, we obtain the entailment probabilities for each hypothesis/verbalization and, finally, we return the most probable hypothesis, that is, we return the CTRE label of the verbalized hypothesis with the highest entailment probability above a threshold. If none overpasses the threshold we return the negative class (OUTREL).

Relation verbalization The verbalization process involves creating one or more templates that represent the temporal relation. The templates include the placeholders {e1} and {e2} that involve the source entity and destination entity of the context. We manually create the templates that verbalize the relations based on the guidelines of the THYME and E3C datasets. In total, we generated between 2-6 templates per label. Table 2 shows the defined templates. Each relation label has entailment and contraction templates that are associated with entailment or contradiction labels when doing the inference with the entailment model.

Entailment model Given a premise and hypothesis, the model returns the probabilities of the hypothesis being entailed by, contradicted to or neutral to the premise. In principle, any model trained on the NLI task can be used. The entailment model used in this work is based on the $\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019) checkpoint (Nie et al., 2020), which has been trained on all SNLI (Bowman et al., 2015), MNLI (Williams, Nangia, and Bowman, 2018), FEVER (Thorne et al., 2018) and ANLI (Nie et al., 2020) datasets[1].

Training The system based on NLI can be used in a zero-shot fashion, but in the ca-

[1]The NLI models used on this work can be downloaded from the HuggingFace Models repository: ynie/roberta-large-snli$_m nli_f ever_a nli_R 1_R 2_R 3 - nli$.

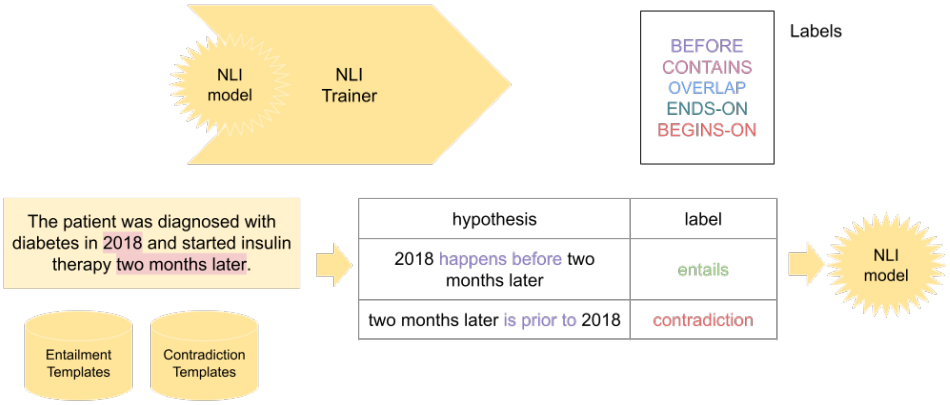

Figura 3: General overview of the training process of the entailment-based CTRE. On the left, input information: the context, the event source (*2018*) and the destination (*diagnosed*). In the middle, hypotheses verbalized using the templates.

se that labeled data exists, we can reformulate the existing dataset into an entailment problem in a similar way that this is done in inference. We generate entailment, neutral and contradiction hypotheses heuristically from the data using the templates defined in Table 2. Given a set of labeled relation examples, we use the following steps to produce entailment pairs for fine-tuning the entailment model. 1) For each positive relation example in the dataset, we generate at least one `entailment` instance by sampling from the corresponding entailment templates of the current CTRE label. That is, we generate one or several premise-hypothesis pairs labeled as entailment. Similarly, 2) for each positive example, a `contradiction` premise-hypothesis is generated by randomly sampling from the corresponding set of contradiction templates (cf. see the contradiction template column in Table 2). Finally, (3) for each negative relation we generate a `neutral` premise-hypothesis pair, sampling a verbalization from either entailment or contradiction templates. We use the reformulated training data for further training the entailment model as shown in Figure 3. For the detection of missing-relations we deploy threshold-based missing-relation detection (Sainz et al., 2021). We apply a threshold value, and if none of the CTRE labels surpasses the threshold, the missing-relation is returned. On the contrary, the relation with the highest entailment probability is returned as the predicted label. In our experimentation, the threshold is estimated as the one that maximizes the macro F1-score in the development set.

5 *Experimental Setup*

This section describes the employed datasets, the evaluation settings, and the baselines and models used in our experiments.

5.1 Datasets

We carried out our evaluation on two datasets with similar annotation guidelines (Wright-Bettner and Palmer, ; Magnini et al., 2022): the THYME corpus (Styler IV et al., 2014) and the E3C corpus (Magnini et al., 2020; Magnini et al., 2021).

Each dataset provides specific entities (EVENTS, TIMEX3) and existing TLINK relations among them across the whole document. Section 5.1.1 and Section 5.1.2 further describe the THYME and EC3 datasets.

5.1.1 THYME corpus

The THYME corpus (Styler IV et al., 2014) consists of de-identified real colon cancer medical records from a hospital annotated following the THYME guides (Wright-Bettner and Palmer,). THYME has both temporal and coreference relations. In our experiments, we use temporal annotations that aim to create a timeline for each patient's notes. The project annotated health records with events that are relevant to the patient's clinical timeline (MRI, surgery, etc.), as well as relevant temporal expressions (such as the date of surgery), and temporal relations between the events (showing whether the surgery came before or after the MRI, for example). THYME TLINKs contain the following relation types: `BEFORE`, `CONTAINS`, `OVERLAP`, `ENDS-ON` and `BEGINS-ON`.

Train split	THYME #Pos	Total	#Docs
1-doc	76	820	1
1 %	187	2087	2
5 %	784	8448	10
10 %	1580	16361	20
25 %	3653	38645	52
FT (100 %)	13840	144141	200

Tabla 3: THYME average Positive relations, total relations and number of documents per 5 random text samples per split (FT = Full Training set).

5.1.2 E3C corpus

As the second dataset, we use the publicly available clinical histories from the E3C corpus (Magnini et al., 2020; Magnini et al., 2021). It is composed of general clinical statements that present the reasons for a clinical visit, the description of physical exams, and the assessment of the patient's situation. The narratives are extracted from multiple medical literature sources, such as PubMed, the Pan African Medical Journal, and the SPACCC corpus, among others. English data was partitioned splitting the set of documents into train, development, and test, sampling 60 %, 20 %, and 20 % of documents at random. Although the corpus contains annotation in four different languages, namely English, Italian, Spanish, French and Basque, the experiments were run only in English.

Train split	E3C EN # Pos	Total	# Docs
1-doc	59	360	1
50 %	884	5000	23
FT	1824	10303	46

Tabla 4: E3C average Positive relations, total relations and number of documents per 5 random text samples per split.

As explained in the guidelines (Magnini et al., 2022), the E3C dataset contains two layers of annotations. We focus on the second layer that includes clinical relevant events, time expressions, and temporal relations according to the THYME standard. We use the gold entity mentions, putting our effort on the detection of TLINKs. E3C contains the following relation types: `BEFORE`, `CONTAINS`, `OVERLAP`, `PERTAINS`, `SIMULTANEOUS`, `ENDS-ON` and `BEGINS-ON`. The labels in the THYME corpus and the E3C corpus are not identical. Notably, the `PERTAINS` label, which represents a relationship between a substance and a measure or quantity, is not a temporal relation but is annotated in the E3C corpus. In contrast, the THYME corpus exclusively annotates temporal relations and does not include the `PERTAINS` label. As our focus is on temporal relations, we have excluded the `PERTAINS` label from the E3C dataset for our analysis. Furthermore, to facilitate the transfer of information between the two corpora, the E3C `SIMULTANEOUS` relation has been mapped to `OVERLAP` although they are not exactly identical.

5.2 Evaluation Settings

Relation and Entity candidates Our evaluation concentrates on predicting the temporal relations between all possible intra-sentence entity pairs. Specifically, we restrict temporal relation extraction to entities occurring within the same sentence, leaving the evaluation of inter-sentence, document-level extraction for future work. Addressing inter-sentence relations differs fundamentally, as it relies heavily on the document's structure and its sections (e.g., chief complaint, personal antecedents, family history, etc.), which is out of the scope of this work. As mentioned in the introduction, the majority of conceivable event-time expression pairings lack any explicit connection in the manual annotation (see Figure 1), we make missing-relations explicit (using `OUTREL` label) to those examples without any CTRE label. This strategy generates an extremely unbalanced dataset towards the `OUTREL` label. Table 3 quantifies the proportions between positive examples and the total number and shows how unbalanced the dataset is.

Evaluation scenarios We created multiple zero- and few-shot settings to measure the robustness of entailment-based approaches compared to the classic supervised models in low-resource scenarios. The **zero-shot** 0 % setting aims to evaluate the default knowledge for temporal reasoning in generic entailment models. The goal of **few-shot** is to measure the adaptability of the models. For that, we create smaller versions of the dataset that simulate low-resource scenarios. In the case of THYME, we sample one document, 1 %, 5 %, 10 %, and 25 % of the documents from the whole corpus. Table 3 shows the sizes of THYME for each training split.

In the case of E3C, we sample one document and 50 % of the documents. The counts are shown in Table 4. We generate 5 random samples for each of the defined splits to measure the variability of the models. The **full-train** (FT) scenario is provided for whole corpus insight.

In addition, similar to zero and few-shot scenarios, we define a **dataset transfer** setting where we measure the transferability of temporal knowledge of models that were trained on temporal reasoning to a new unseen dataset. Continual fine-tuning settings seek to measure the adaptability of entailment models trained on a different temporal relation extraction setting.

5.3 Evaluation metrics

We have used the standard F1-Score, which is a common metric on IE tasks. In particular, we report the macro average of F1-Score over the positive labels, as standard practice in information extraction. The macro average captures better the overall performance for this particular setting, where around $80 - 90$ % of examples are missing-relations (`OUTREL`). Focusing on the micro average we would ignore the effectiveness of correctly predicting less frequent labels. Table 5 presents the label distribution of the evaluation set in THYME and E3C, highlighting the challenges of this particular setting. Note that the reported results represent the average of 5 random samples of each training split.

Label	E3C	THYME
OUTREL	1847 (81.29 %)	61066 (89.45 %)
CONTAINS	242 (10.65 %)	4573 (6.70 %)
BEFORE	53 (2.33 %)	970 (1.42 %)
OVERLAP	75 (3.30 %)	1168 (1.71 %)
ENDS-ON	19 (0.84 %)	134 (0.20 %)
BEGINS-ON	36 (1.58 %)	361 (0.53 %)

Tabla 5: Label distribution of the evaluation sets in E3C and THYME corpora.

5.4 Baselines and Models

Our primary baseline for comparison is our re-implementation of the Entity Marker input representation (EM) (Soares et al., 2019). Entity markers help relation extraction models by embedding structural and semantic information about entity pairs, improving global feature learning. Additionally, a classification layer is added to predict the actual temporal relation between entities based on the entity marker vectors. The EM strategy is evaluated on the same few-shot splits as our system, enabling a direct head-to-head comparison. EM, a state-of-the-art model (Zhou and Chen, 2022), leverages RoBERTa$_{\text{LARGE}}$ as its pre-trained models and the relation representation obtained through entity markers. The model is a general-domain model. Since the authors did not find a medical domain-specific NLI model, we also avoided using a medical domain-specific RoBERTa for comparison purposes. In addition, we implemented two simple baselines, that aim to contrast the obtained results against randomness and the simplest classifier. On one hand, we compare our results with the **majority** baseline, which always predicts the most frequent positive label (`CONTAINS`). On the other hand, the **random** baseline assigns classes uniformly (ignoring the class distribution).

Parameter	NLI Value	EM Value
Learning Rate	2e-5	1e-5
Batch size	8	4
Max. epochs	7	20
weight-decay	0.01	0.01

Tabla 6: Hyperparameter values for NLI-based and EM approaches.

6 Results

6.1 Main Results

Once we checked that our EM implementation obtained similar results to those of the SOTA (Lin et al., 2021b), we evaluated the described experiments. Table 7 shows the results for in-domain experiments carried out in THYME and E3C, separately. It summarizes the performances in zero-shot (0 %), few-shot (from 1-DOC to 25 %) and full training (FT) settings. The experiments show that the entailment-based approach (NLI) consistently outperforms EM when few annotated data is available. For example, up to 10 % of the whole set of annotations in THYME. Once there is enough annotated data EM becomes more competitive than NLI, at least in THYME. Hypothetically, we would need to annotate around 50 THYME documents (circa 38,000 training examples) to train a competitive supervised classifier.

	THYME							E3C EN			
Model	0 %	1DOC	1 %	5 %	10 %	25 %	FT	0 %	1DOC	50 %	FT
Majority	-	2.5	2.5	2.5	2.5	2.5	2.5	3.8	3.8	3.8	3.8
Random	**7.5**	7.5	7.5	7.5	7.5	7.5	7.5	**9.1**	**9.1**	9.1	9.1
EM	-	7.7$_{\pm 2,6}$	13.5$_{\pm 4,3}$	20.2$_{\pm 9,1}$	29.8$_{\pm 2,9}$	**35.1**$_{\pm 1,2}$	**40.5**	-	3.0$_{\pm 2,9}$	8.8$_{\pm 5,0}$	16.1$_{\pm 1,6}$
NLI	5.4	**11.1**$_{\pm 2.7}$	**14.8**$_{\pm 3.3}$	**25.0**$_{\pm 3.0}$	**30.5**$_{\pm 2.5}$	34.5$_{\pm 1.9}$	36.5	8.2	8.7$_{\pm 2.3}$	**21.7**$_{\pm 2.5}$	**21.4**$_{\pm 1.2}$

Tabla 7: Main results (Macro average F1-Score) comparing the performance of supervised classifier (EM) and entailment-based approach (NLI), along with the random and majority class baselines.

THYME → E3C EN							
Model	0 %	1DOC	1 %	5 %	10 %	25 %	FT
Majority	-	3.8	3.8	3.8	3.8	3.8	3.8
Random	9.1	9.1	9.1	9.1	9.1	9.1	9.1
EM	-	2.9$_{\pm 3,0}$	6.5$_{\pm 2,9}$	10.3$_{\pm 5,3}$	13.8$_{\pm 1,8}$	16.3$_{\pm 2,6}$	**19.2**
NLI	8.2	**9.2**$_{\pm 4,3}$	**10.0**$_{\pm 2,3}$	**12.9**$_{\pm 2,7}$	**16.9**$_{\pm 3,2}$	**16.5**$_{\pm 1,3}$	17.1

Tabla 8: Zero Shot Transfer Learning (Macro average of F1-Score) from different THYME in domain scenarios (0 %, 1doc, 1 %, 5 %, 10 %, 25 % and FT THYME proportions) to E3C EN via pretrained checkpoints.

In E3C results are alike, NLI consistently outperforms EM in all few-shot and FT settings. It is worth noting that the EM results in E3C are lower than expected in the 50 % setting. A deeper look revealed problems of the models to generalize from training.

In the zero-shot setting (0 % columns in the tables), the entailment model cannot outperform the random baseline in any dataset used in the evaluation. This is indicative that generic entailment models do not have the knowledge to infer temporal relations correctly and have to be acquired from other resources in addition to the NLI datasets.

6.2 Dataset Transfer Results

Table 8 shows the results of training a model in THYME and testing it in E3C. The comparison of EM with NLI shows a similar trend as was described in the previous section: when few annotated data are available, the entailment model has better transferability across datasets (from THYME to E3C). That is, NLI requires less training data for effective cross-dataset learning.

A closer look at the results reveals that, overall, the training in THYME is quite effective when deploying the models (both EM and NLI) in the E3C dataset. Note that, using a 10 % of THYME for training the NLI model is equivalent to training EM on the whole E3C training set. Training an NLI model in 10 % of THYME attains 16.9 of F1-Score while training EM on the full training set reaches 16.2 of F1-Score (cf. Table 7).

6.3 Continual Fine-tuning Results

Tables 9 and 10 show the results of continued fine-tuning of the entailment and EM models. That is, we evaluate the transferability of the models previously trained on a different dataset (THYME) to a new training dataset (E3C). The tables show the amount of training used from the E3C dataset. Note that in this setting the models are evaluated always in E3C. We selected the best checkpoints from THYME 25 % and FT splits as the starting point of the continual setting.

The results show that in general entailment models show better capabilities for transferring temporal knowledge compared to the EM models. As expected, EM struggles to combine acquired knowledge from the old and new training datasets, and only improves when sufficient new training data is available (cf. Table 9). On the contrary, the entailment model (NLI) obtains consistent improvements across the training splits in E3C, with the exception of the 1DOC setting.

The amount of training in THYME affects the adaptation of the models. The larger the amount of training in THYME the bigger the difficulties to be adapted to E3C. Using full training in THYME, the model is not able to improve the results in E3C and the F1-Score decreases from 19.2 to 16.8. Strong fine-tuning in THYME affects to a lesser extent the entailment model, where it only lo-

THYME 25 % → E3C EN				
Model	0 %	1DOC	50 %	FT
Majority	-	3.8	3.8	3.8
Random	9.1	9.1	9.1	9.1
EM_{+e3c}	15.9	**16.8**$\pm$4.5	14.4$\pm$4.0	20.7$\pm$4.1
NLI_{+e3c}	**18.3**	8.7$\pm$2,3	**21.5**$\pm$2,7	**21.4**$\pm$2,2

Tabla 9: Continual pretrain (MacroF1-Score) over best THYME 25 % checkpoint.

THYME FT → E3C EN				
Model	0 %	1DOC	50 %	FT
Majority	-	3.8	3.8	3.8
Random	9.1	9.1	9.1	9.1
EM_{+e3c}	**19.2**	**16.0**$\pm$2.8	13.6$\pm$5.2	16.8$\pm$7.9
NLI_{+e3c}	17.1	14.2$\pm$3.3	**19.7**$\pm$1.1	**19.0**$\pm$1.8

Tabla 10: Continual pretrain (MacroF1-Score) over best THYME FT checkpoint.

ses 1.4 points. Finally, experiments show that the use of a single document is not sufficient to adapt the models to new scenarios. For the entailment model, it attains worse results when a single document is added from the E3C training (a decrease from 18.3 to 8.7).

7 Discussion

7.1 Transfer Learning

The transfer learning experiments reveal that reformulating the task as entailment is more robust than using EM. By abstracting information to a higher semantic level, the entailment approach achieves better generalization, particularly across different labels and scenarios, as shown when comparing Tables 11 and 12. Notably, as expected, `OVERLAP` exhibits higher variability among the labels. It is worth recalling that we equate the E3C's `SIMULTANEOUS` label with THYME's `OVERLAP`, though they are not entirely equivalent. Despite this, the entailment-based approach demonstrates superior generalization compared to the BERT-based ER classifier across most labels. The only exception is `BEGINS-ON`, which shows comparable performance in both approaches.

Related to the transfer generalization it is important to mention that although both corpora belong to the medical domain, THYME holds real admission records while E3C contains medical literature articles. As (Lin et al., 2021b) pointed out, the language of biomedical literature is different from the clinical notes found in electronic medical records. The NLI ability to better transfer information across both datasets suggests that the CTRE task recast improves semantic abstraction achieving better temporal reasoning.

Label	E3C	Transfer	CONT. 25 %	CONT. FT %
OUTREL	68.69	83.62	70.20	76.51
CONTAINS	35.86	33.42	33.78	33.73
BEFORE	23.76	26.37	23.61	29.14
OVERLAP	13.62	04.30	14.42	12.90
ENDS-ON	20.25	22.22	22.22	11.32
BEGINS-ON	19.60	09.52	30.00	21.05

Tabla 11: NLI E3C FT vs NLI THYME → E3C EN FT per class results. CONT. refers to CONTINUAL pretrain.

Label	E3C	Transfer	CONT. 25 %	CONT. FT %
OUTREL	88.83	83.62	87.26	88.03
CONTAINS	17.94	33.42	25.48	31.05
BEFORE	25.88	26.37	30.63	20.89
OVERLAP	04.81	04.30	0	02.46
ENDS-ON	17.39	22.22	22.22	16.00
BEGINS-ON	25.92	09.52	34.48	32.25

Tabla 12: EM E3C FT vs EM THYME → E3C EN FT per class. CONT. refers to CONTINUAL pretrain.

8 Conclusions

We demonstrated that an entailment-based reformulation effectively models complex temporal relationships in the medical domain. Our findings indicate that this approach outperforms BERT-based relation extraction methods in low data scenarios, aligning with the conclusions of (Sainz et al., 2021) for factual relations. Furthermore, we showed that entailment-based methods facilitate effective transfer between datasets, particularly in low-data regimes, employing continual pretraining. This involves selecting the best checkpoint from a subset of the source dataset and continuing pretraining on a subset of the target dataset. Our results suggest that while BERT-based RE classifiers perform better using ample data, entailment-based reformulations excel in low-data settings, mitigating majority-class bias. Finally, although generic entailment models exhibit limited temporal knowledge, such knowledge can be effectively acquired from even the most scarce existing data sources.

Limitations and Future Work

This work utilizes general-domain NLI models, which may not be optimal for the clinical domain. It focuses on same-language transfer learning, while cross-language transfer remains a future goal, particularly for underrepresented languages like Spanish, French, and Italian. The study demonstrates that LLMs can learn temporal reasoning and transfer knowledge across tasks, raising the question of whether unsupervised Silver corpora generated through data augmentation could enhance this ability.

Future research will explore the role of narrative containers (e.g., the CONTAINS label) in constructing fully connected knowledge graphs from clinical notes. Additionally, the study aims to compare NLI's effectiveness in mitigating the majority class bias with other intermediate tasks, such as QA. Lastly, the authors plan to investigate the integration of temporal knowledge from multiple datasets (e.g., MACCROBAT, eHealthKD, THYME, E3C) to enhance LLMs' temporal reasoning capabilities through continual training.

Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation (MCI/AEI/FEDER, UE, EDHER-MED PID2022-136522OB-C22), the Basque Government (IXA excellence research group IT1570-22, ELkartek CL4LANG, KK-2023/00094 and IKER-GAITU 11:4711:23:410:23/0808), LUMINOUS (European Union Horizon Europe HORIZON-CL4-2023HUMAN-01-21-101135724), DeepKnowledge (Spanish Ministry MCIN PID2021127777OB-C21 AEI/10.13039/501100011033. This work has been done under the LINGUATEC-IA project (EFA104/01) which is co-financed at 65 % by the European Regional Development Fund (ERDF) through the Interreg V-A Spain-France-Andorra Programme (POCTEFA 2021-2027).

References

Alfattni, G., N. Peek, and G. Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, 108:103488.

Allen, J. F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Baldini Soares, L., N. FitzGerald, J. Ling, and T. Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July. Association for Computational Linguistics.

Baucells, I., B. Calvo, M. Villegas, and O. L. de Lacalle. 2023. Entailment-based task transfer for catalan text classification in small data regimes. *Procesamiento del Lenguaje Natural*, 71(0):165–177.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning. 2015. A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Bui, A. A. T., D. R. Aberle, and H. Kangarloo. 2007. Timeline: Visualizing integrated patient records. *IEEE Transactions on Information Technology in Biomedicine*, 11:462–473.

Caron, A., E. Chazard, J. Muller, R. Perichon, L. Ferret, V. Koutkias, R. Beuscart, J.-B. Beuscart, and G. Ficheur. 2017. Itcares: an interactive tool for case-crossover analyses of electronic medical records for patient safety. *J Am Med Inform Assoc*, 24:323–330, 03/2017.

Dagan, I., O. Glickman, and B. Magnini. 2006. The pascal recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dalianis, H. 2018. *Clinical text mining: Secondary use of electronic patient records.* Springer Nature.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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, June. Association for Computational Linguistics.

Hirsch, J., J. Tanenbaum, S. Gorman, C. Liu, E. Schmitz, D. Hashorva, A. Ervits, D. Vawdrey, M. Sturm, and N. Elhadad. 2014. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association : JAMIA*, 22, 10.

Johnson, A. E. W., M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford. 2016. Machine learning and decision support in critical care. *Proceedings of the IEEE*, 104(2):444–466.

Knez, T. and S. Žitnik. 2024. Multimodal learning for temporal relation extraction in clinical texts. *Journal of the American Medical Informatics Association*, 31(6):1380–1387, 03.

Lin, C., T. Miller, D. Dligach, S. Bethard, and G. Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Lin, C., T. Miller, D. Dligach, S. Bethard, and G. Savova. 2021a. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, editors, *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online, June. Association for Computational Linguistics.

Lin, C., T. Miller, D. Dligach, S. Bethard, and G. Savova. 2021b. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, editors, *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online, June. Association for Computational Linguistics.

Lin, Y., K. Lu, S. Yu, T. Cai, and M. Zitnik. 2023. Multimodal learning on graphs for disease relation extraction. *Journal of Biomedical Informatics*, 143:104415.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In *CLiC-it*.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2021. The e3c project: European clinical case corpus. *Language*, 1(L2):L3.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2022. E3c annotation guidelines. Accessed: 2022-03-04.

Nie, Y., A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July. Association for Computational Linguistics.

Ning, Q., Z. Feng, and D. Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ning, Q., S. Subramanian, and D. Roth. 2019. An improved neural baseline for temporal relation extraction. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro-*

cessing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6203–6209, Hong Kong, China, November. Association for Computational Linguistics.

Obamuyide, A. and A. Vlachos. 2018. Zero-shot relation classification as textual entailment. In J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, editors, *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium, November. Association for Computational Linguistics.

Olex, A. L. and B. T. McInnes. 2021. Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be. *Journal of biomedical informatics*, page 103784.

Poliak, A., A. Haldar, R. Rudinger, J. E. Hu, E. Pavlick, A. S. White, and B. Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium, October-November. Association for Computational Linguistics.

Sainz, O., I. Gonzalez-Dios, O. Lopez de Lacalle, B. Min, and E. Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multisource learning. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States, July. Association for Computational Linguistics.

Sainz, O., O. Lopez de Lacalle, G. Labaka, A. Barrena, and E. Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Soares, L. B., N. Fitzgerald, J. Ling, and T. Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Styler IV, W. F., S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, and J. Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Thorne, J., A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

Uppal, S., V. Gupta, A. Swaminathan, H. Zhang, D. Mahata, R. Gosangi, R. R. Shah, and A. Stent. 2020. Two-step classification using recasted data for low resource settings. In K.-F. Wong, K. Knight, and H. Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 706–719, Suzhou, China, December. Association for Computational Linguistics.

van der Linden, S., J. J. van Wijk, and M. Funk. 2021. Multiple scale visualization of electronic health records to support finding medical narratives. In *VCBM*.

Vashishtha, S., A. Poliak, Y. K. Lal, B. Van Durme, and A. S. White. 2020. Temporal reasoning in natural language inference. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online, November. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, J., X. Le, X. Peng, and C. Chen. 2023. Adaptive hinge balance loss for document-level relation extraction. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3872–3878, Singapore, December. Association for Computational Linguistics.

Wang, Q., T. Mazor, T. A. Harbig, E. Cerami, and N. Gehlenborg. 2022. Threadstates: State-based visual analysis of disease progression. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):238–247.

Williams, A., N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

Wright-Bettner, K. and M. Palmer. Synthesizes and expands on the thyme annotation guidelines, clinical coreference annotation guidelines, and richer event description (red) annotation guidelines. Accessed: 2024-12-01.

Zhou, W. and M. Chen. 2022. An improved baseline for sentence-level relation extraction. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only, November. Association for Computational Linguistics.

Zhou, Y., Y. Yan, R. Han, J. H. Caufield, K.-W. Chang, Y. Sun, P. Ping, and W. Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *AAAI*, pages 14647–14655. AAAI Press.