Temporal Relation Classification: An XAI Perspective

Sofia Elena Terenziani
IT University of Copenhagen
seterenziani@gmail.com

Abstract

Temporal annotations are used to identify and mark up temporal information, offering definition into how it is expressed through linguistic properties in text. This study investigates various discriminative pre-trained language models of differing sizes on a temporal relation classification task. We define valid reasoning strategies based on the linguistic principles that guide commonly used temporal annotations. Using a combination of saliency-based and counterfactual explanations, we examine if the models' decisions are in line with the strategies. Our findings suggest that the selected models do not rely on the expected linguistic cues for processing temporal information effectively [1].

1 Introduction

Temporal information processig is a fundamental aspect of natural language and is essential for NLP applications including question answering (Chen et al., 2021; Ko et al., 2023), text summarization (Daiya, 2020), and information retrieval (Gade and Jetcheva, 2024). Transformer-based pre-trained language models have shown impressive performance in such tasks (Xiong et al., 2024; Ko et al., 2023; Tai, 2024; Shi et al., 2023). Yet, their interpretation of time diverges from human interpretation (Callender, 2011), making it challenging to evaluate their temporal processing, and whether they indeed interpret the temporal information as expected (Qiu et al., 2023; Jain et al., 2023).

While temporal benchmarks (Tan et al., 2023a; Zhou et al., 2019; Ning et al., 2020; Zhou et al., 2021) have been extensively developed, performance metrics alone do not reveal the underlying mechanisms or explain how conclusions are reached (Chakraborty et al., 2017). This study contributes a methodology and an evaluation dataset for evaluating NLP models on temporal relation classification. We define valid reasoning strategies, and use a combination of saliency-based and example-based explainability methods to assess whether a model follows these strategies when making decisions.

Our framework extends the work introduced by Ray Choudhury et al. (2022). We explore discriminative models of varying sizes to determine if larger models, trained more extensively on more data, are also more likely to base their decisions on valid information retrieval processes. Our findings suggest that while larger models show better performance on the task, they frequently deviate from expected reasoning strategies. These results align with broader concerns about the reliability of current popular benchmarks, where high accuracy can mask a reliance on shortcuts or spurious correlations. We discuss the limitations of this framework, together with the opportunities and challenges of extending it to generative models.

2 Related Work

Temporal Relation Classification. Temporal relation classification (TRC) was first introduced in TempEval-3 (UzZaman et al., 2013) and gained popularity with dedicated corpora and annotations for temporal information processing. Modern TRC methods predominantly use discriminative pre-trained language models, to generate robust contextual representations for pairs of event mentions (Yang et al., 2019; Lin et al., 2019). Further advancements include graph-based methods (Mathur et al., 2021; Zhang et al., 2022; Zhou et al., 2022) and prompt and masking techniques (Han et al., 2021; Yang et al., 2024). Despite the recent surge in the generative models, they still underperform compared to fine-tuned smaller mod-

[1] https://github.com/sofitere/TRC-XAI

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 714–728
March 3-4, 2025 ©2025 University of Tartu Library

	Reasoning Step	**Relevant Features**
Context: Leon won the marathon years after he underwent surgery in 2011.	Identify temporal information	Expression: *years, 2011* Preposition: *after*
	Map temporal information to event	*underwent* := *2011* *won* := (*years, after*)
Relation: ⟨won, **?**, underwent⟩	Determine temporal relationship	*won* := *year after 2011* ⟨won, **AFTER**, underwent⟩

Table 1: Valid reasoning steps for determining the temporal relation between a given event pair.

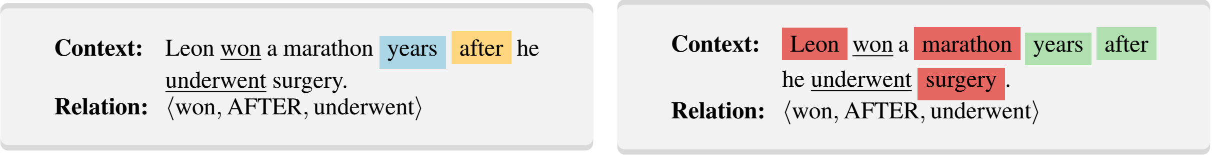
els (Roccabruna et al., 2024; Yuan et al., 2023).

Temporal Annotation. TimeML (Mani et al., 2006) remains the most widespread format for temporal annotation, and it became the basis for ISO standard (Pustejovsky et al., 2010). TimeML includes conventions to identify and describe temporal elements in text, including temporal expressions (`TIMEX`), events, temporal relations (`T-LINKS`), signals (`SIGNAL`), and relation types. TimeBank corpus (Pustejovsky et al., 2003) has been re-annotated in several projects to increase the density of T-LINKs (Verhagen et al., 2007; Rogers et al., 2022; Naik et al., 2019) and improve its consistency. Its texts have been utilized in subsequent projects providing additional annotation in other formats, including MATRES (Ning et al., 2018).

Benchmarks. Benchmarks for temporal processing vary widely in format and scope. TimeQA (Chen et al., 2021) and Tempreason (Tan et al., 2023b) focus on temporal question answering, Torque (Ning et al., 2020) on temporal reading comprehension, adopting question/answering as format, and MCTACO (Zhou et al., 2019) on temporal commonsense reasoning, adopting multiple-choice as format. Commonly used benchmarks have shown some limitations, also here ranging from task and scope. Temporal question-answering (QA) benchmarks tend to be biased in their coverage of time spans and question types, leading to models performing well due to format biases rather than actual language processing skills (Tan et al., 2023c). Additionally, benchmarks with focus on temporal expressions, such as numeric years, have shown to not represent the full range of diversity of temporal expressions (Qin et al., 2021). Benchmarks for reading comprehension often assume that performing well necessitates engaging with cognitive processes of language understanding (Sugawara et al., 2019; Weston et al., 2015), implying that higher scores reflect advances in general language processing (Ray Choudhury et al., 2022). Performance on benchmarks alone, while useful, does not necessarily tell us whether the model is right for the right reasons; if it is not, the benchmark results may be misleading and not generalize to other data (Dehghani et al., 2021; Bowman and Dahl, 2021).

Explainability. Explainability methods can account for some of the limitations of the current benchmarks by highlighting what information the model relies on, or where it fails to perform. They can thus provide means to check to what degree the models are reliable, i.e. they perform correctly and consistently for the right reasons (McCoy et al., 2019; Christianson, 2016). For this line of research, local and post-hoc methods have been used to evaluate pre-trained language models on tasks that demand specific linguistic skills. Ray Choudhury et al. (2022) apply a combination of these methods to analyze and evaluate models on two linguistic skills required for a reliable reading comprehension system, finding that models use shortcuts rather than valid inference strategies. In the context of LLMs, explainability methods are both important and challenging. Research efforts are also put into examining the utility (González et al., 2021), interpretability (González et al., 2021; Schuff et al., 2022) and reliability (Harbecke and Alt, 2020; Spreitzer et al., 2022; Rahimi and Jain, 2022) of explainability methods.

Contribution. To date, relevant NLP work on temporal processing has focused on modeling, benchmarks and annotation schemes. The use of explainability methods to explore how models handle temporal data is largely unexplored. To the best of our knowledge, this is the first study to apply saliency-based and example-based interpretability methods to assess whether models rely on the expected reasoning patterns for temporal relation classification. We evaluate the validity



Context: Leon won a marathon years after he underwent surgery.
Relation: ⟨won, AFTER, underwent⟩

(a) Example of temporal annotation: TIMEX (blue), and SIGNAL (orange). These are temporal elements relevant for expressing the relationship between 'won' and 'underwent'.

Context: Leon won a marathon years after he underwent surgery .
Relation: ⟨won, AFTER, underwent⟩

(b) Example of token partitioning into positive (green) and negative (red). The models are expected to rely more on the tokens in the 'positive' set.

Figure 1: A sample question from the MATRES (Ning et al., 2018) dataset. A model is asked to predict the temporal relationship between winning a marathon and having brain surgery. Token partitioning is delivered from the features defined as relevant for determining the temporal relation between two events.

of these methods (via examining their alignment), and discuss the challenges of evaluating the latest generative models on temporal relation classification.

3 Defining Success Criteria for TRC

Evaluating whether models follow expected reasoning involves testing if their decision-making process are based on valid information retrieval and inference strategies rather than superficial patterns in the data. Ray Choudhury et al. (2022) defines three success criteria for NLP systems: a system must (1) accurately perform on a specific task, (2) rely on information deemed pertinent to the task, and (3) maintain consistency under distribution shifts. We evaluate a model's performance in TRC against these criteria. We first define the expected reasoning processes (§ 3.1). We then assess the model's adherence to these reasoning steps by verifying its reliance on valid information (§ 4.7), and by evaluating its performance consistency across variations in data distribution (§ 4.6).

3.1 What reasoning should a model perform?

To correctly extract and classify temporal relations, a model must identify linguistic features that express temporal information, map these features to the events they describe or modify, and use this information to deduce the temporal relationship between the pair of events. We define these as valid reasoning steps[2] (see Table 1 for an example). Temporal annotation schemes and guidelines can be used to clarify which linguistic features are essential for identifying the temporal relation between an event pair. We focus on two types of annotations from the TimeML guidelines (Mani et al., 2006):

- `TIMEX3` tags are utilized for annotating explicit temporal expressions within text. These expressions can be absolute (”December 2025”, ”5PM”) or relative (”Mondays”, ”monthly”). They serve to anchor events to specific times or durations.
- `SIGNAL` tags mark words or phrases that cue the relationships between two entities (e.g. timex to event, timex to timex, event to event). Common linguistic features are adverbs (”again”, ”late”, eventually”) detailing the timing of events, conjunctions (”before”, ”since”, ”while”) relating events to each other and subordinate conjunctions (”because”, ”if”, ”therefore”) expressing conditional or causal relationships. These features indicate the sequence or structure of events, showing their interactions over time.

Essentially, while `TIMEX3` tags are used to identify temporal entities, `SIGNAL` annotations establish the links between these entities within the text. Together, they provide the foundational information necessary to understand the temporal relationships among events in texts.

3.2 What reasoning does a model perform?

Having established the reasoning processes a model should follow, the next step is to assess whether a specific model adheres to these. Ray Choudhury et al. (2022) uses a combination of example-based and saliency-based interpretability methods. These methods are categorized as local and post-hoc (Molnar, 2022): they focus on individual instances and they are applied after model

[2] We recognize that this represents only the minimal information on which models (or humans) might rely. For the example shown in Table 1, if the context includes details about Leon breaking his leg, this information could reasonably influence the understanding of Leon's chances of winning the marathon. Nonetheless, the minimally necessary information in the immediate context would still be salient, and it is a reasonable expectation that either models or humans should rely on it.

	Purpose	# Docs	#Events	#TLinks
TimeBank	Training	162	6.6k	6.5k
Aquaint	Training	73	4,3k	6.4k
Platinum	Validation	20	748	837
Total		275	6k	13.5k

Table 2: Summary of purpose and statistics of the MATRES (Ning et al., 2018) dataset subsets.

Label	#	%
BEFORE	6.886	50%
AFTER	4.576	34%
VAGUE	1.644	12%
EQUAL	471	4%

Table 3: Label distribution in the MATRES (Ning et al., 2018) dataset.

has been trained.

Saliency-based Methods. Saliency-based methods are a family of methods that offer feature-centered explanations (Molnar, 2022; Ding and Koehn, 2021a). These methods offer different ways of computing a score for each token, indicating how individual features (token) affect a model's decision. By comparing the saliency scores to a predefined partition of tokens, these explanations can be used to determine whether a model is relying on the right information for correct predictions. Following Ray Choudhury et al. (2022), we define a partition of the token space as: tokens a model should find important (positive), and tokens a model should not find important (negative) (§ 4.5). If saliency scores show that a model consistently has higher scores on the positive compared to the negative partition of tokens, it suggests that the model focuses on the 'right' information.

Counterfactual Explanations. Counterfactual explanations offer data-centred explanations by analyzing how changes in the input data can lead to different model predictions (Molnar, 2022). By changing parts of the input with alternative valid tokens that would change the type of temporal relation, these explanations can help determine if a model is relying on the expected reasoning strategies (§ 4.6). If a model predicts the correct temporal relationship for both original and altered inputs, it suggests that the model consistently relies on the correct information.

Explanation Alignment. For a model to demonstrate valid reasoning, both saliency and counterfactual explanations must align across many instances, suggesting that a model consistently relies on the right information for accurate predictions.

4 Methodology

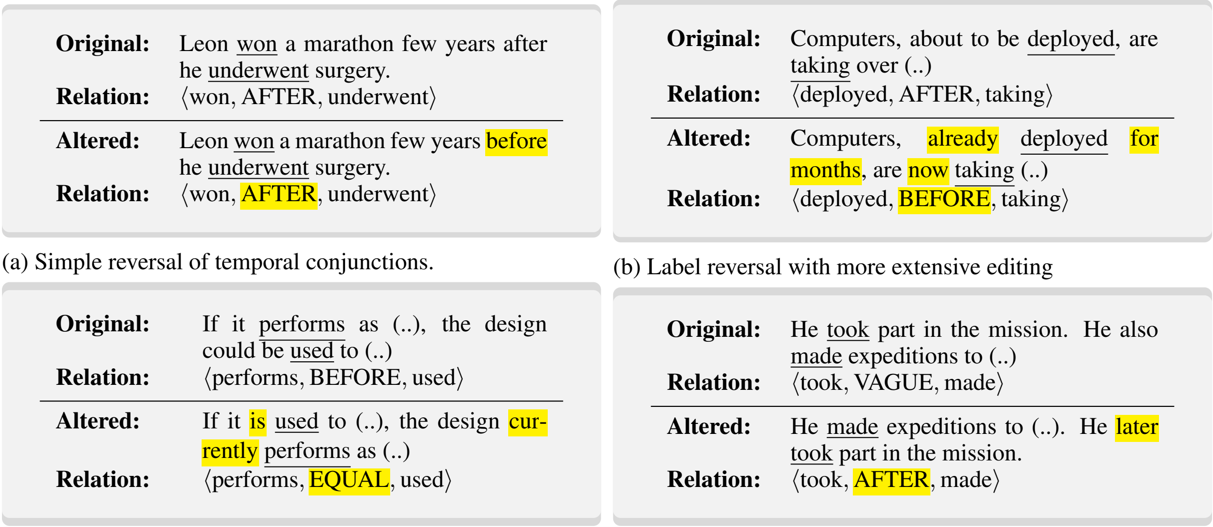
4.1 Data

Our experiments are conducted on the MATRES dataset (Ning et al., 2018). In total, MATRES includes 275 news articles from TempEval3 (UzZaman et al., 2013), annotated for temporal relations between pair of events. For experimental consistency, we follow the original split for training and evaluation (Ning et al., 2019), as shown in Table 2. MATRES is annotated for four different temporal relation classes. The label distribution is shown in Table 3.

4.2 Models

We experiment with transformer-based encoders of different sizes from three families: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and LUKE (Yamada et al., 2020). BERT and RoBERTa are the classical models to use for this task; they share a similar architecture but differ in pre-training scope and optimization (with RoBERTa also receiving more extensive training, but without optimization for the next-sentence-prediction task). They have been used extensively for temporal relation classification (Liu et al., 2019).

We also add LUKE (Yamada et al., 2020): the model enhancing the RoBERTa framework with entity-aware self-attention, improving contextual understanding. Since entities are crucial to temporal relation classification (e.g. for recognizing dates and events), this model could be expected to improve on base BERT/RoBERTa. For all models, we experiment with 'base' and 'large' versions. For some cases, larger models have shown to generalise better (Zhong et al., 2021; Desai and Durrett, 2020). Part of this project is set to investigate whether they are also more likely to rely on the right information. We focus on discriminative models, as they are known for their robust performance in TRC (§ 2). While incorporating gener-



Original: Leon won a marathon few years after he underwent surgery.
Relation: ⟨won, AFTER, underwent⟩

Altered: Leon won a marathon few years before he underwent surgery.
Relation: ⟨won, AFTER, underwent⟩

(a) Simple reversal of temporal conjunctions.

Original: Computers, about to be deployed, are taking over (..)
Relation: ⟨deployed, AFTER, taking⟩

Altered: Computers, already deployed for months, are now taking (..)
Relation: ⟨deployed, BEFORE, taking⟩

(b) Label reversal with more extensive editing

Original: If it performs as (..), the design could be used to (..)
Relation: ⟨performs, BEFORE, used⟩

Altered: If it is used to (..), the design currently performs as (..)
Relation: ⟨performs, EQUAL, used⟩

(c) Changing a conditional relationship

Original: He took part in the mission. He also made expeditions to (..)
Relation: ⟨took, VAGUE, made⟩

Altered: He made expeditions to (..). He later took part in the mission.
Relation: ⟨took, AFTER, made⟩

(d) Sentence reordering

Figure 2: Examples of counterfactual alterations changing the original temporal relation label, with altered tokens highlighted in yellow.

ative models could be insightful, their limitations within this framework are addressed in Section 7.

4.3 Fine-Tuning

Each encoder is fine-tuned for TRC using the tokenization strategy proposed by Yanko et al. (2023) and Baldini Soares et al. (2019). The strategy consists in explicitly marking the boundaries of each event in an input sentence with special tokens. We define these as `[a1]`, `[/a1]`, `[a2]`, `[/a2]` and process each input sentence as following:

Leon `[a1]` won `[/a1]` a marathon years after he `[a2]` underwent `[/a2]` surgery.

When a given input is processed by each encoder, the embeddings of the special tokens are adjusted based on surrounding tokens. This results in a context-specific representation for each event. We concatenate the embedding vectors of the special tokens and use them for classification by feeding them into a linear layer on top of each encoder. All code to reproduce our results, including hyperparameters, is included with the submission and will be made public upon acceptance of the paper.

4.4 Evaluation Metrics

We evaluate each encoder using standard evaluation metrics for classification: F1 and exact-match. Given the significant class imbalance in the MATRES dataset (see Table 3), the F1-score is particularly important. We report both weighted and macro-average F1-score. Although exact-match is less reliable for imbalanced datasets, we include it for its straightforward interpretability.

4.5 Token partition

We previously defined linguistic features essential for expressing the temporal relationships between events (§ 3.1). Token partitioning is guided by this definition. The positive token partition is defined as all individual tokens that express or clarify the temporal relationship between two events, such as temporal expressions, prepositions, conjunctions, and verbs demonstrating tense and aspect. The negative token partition is defined as tokens that are not part of the positive partition and do not match the relevant tokens for the event pair, deemed irrelevant for expressing the temporal relationship. Figure 1 shows the relevant tokens for an instance, and how these define the partition of tokens.

4.6 Counterfactual Explanations

Counterfactual explanations are crafted from 300 instances randomly selected from the validation dataset, with minimal modifications to the original input. The queried event pair to the temporal relation is kept intact[3], and changes are limited

[3] Alterations often involve reversing verb tenses. Since event pairs are defined by the verb's base form and English verb tenses are structured flexibly, most instances can be altered without changing the original event pair. However, shifting to perfect tenses (e.g., "will finish," "had finished"), which useful to indicate completed events isn't always possible.

	F1 M/avg	F1 W/avg	EM
LUKE large	0.54	0.70	0.70
LUKE base	0.55	0.67	0.68
RoBERTa large	0.58	0.70	0.72
RoBERTa base	0.56	0.69	0.69
BERT large-uncased	0.58	0.69	0.69
BERT base-uncased	0.52	0.66	0.66

Table 4: Performance of different models on the MATRES (Ning et al., 2018) dataset.

	Original	Counterfactual
LUKE large	0.66	0.45
LUKE base	0.60	0.43
RoBERTa large	0.67	0.43
RoBERTa base	0.63	0.41
BERT large-uncased	0.62	0.40
BERT base-uncased	0.61	0.44

Table 5: Performance on counterfactual vs. original instances (measured as F1 W/avg).

to the surrounding context. The alteration process involves a two-stage approach: (a) identifying the positive partition of tokens (§ 4.5), likely to impact predictions significantly, and (b) modifying these to change the temporal relationship.

We made alterations of four types, presented in Table 2. About 67% of instances are altered by reversing temporal conjunctions (e.g., modifying "before" or "after"), or adding modifiers or temporal expressions. This strategy is often applied to alter BEFORE-AFTER relationships, aligning with the dataset's label distribution, where these are the most common labels. Less frequent methods like reversing phrase order ($\approx 12\%$) and changing conditional relationships ($\approx 21\%$) targeted the rarer EQUAL and VAGUE labels.

4.7 Saliency Scores

We obtain saliency scores from two different methods: Occlusion and Integrated Gradients (IG).

Occlusion (DeYoung et al., 2020) is a perturbation-based method. It works by systematically replacing the input token with a baseline token and observing the changes in the model's output probabilities. The occlusion score for each token represents the change in the model's output probability when the token is occluded. We select `[MASK]` as the baseline token to represent the absence of a specific feature. By replacing each token one at a time with `[MASK]`, we remove the specific information provided by that specific token and observe how its absence affects the model's output.

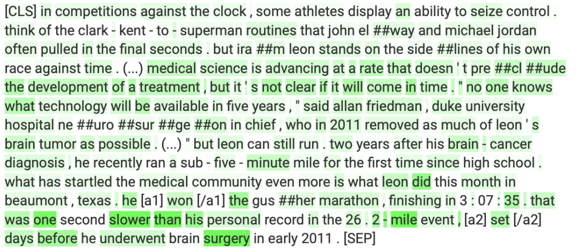
Integrated gradient (Sundararajan et al., 2017; Molnar, 2022) is a gradient-based method. This family of methods work by quantifying how much each token in an input contributes to the gradient being propagated downstream. Tokens that have larger impact on the output will impact the gradient more, and are considered more influential. IG work by comparing the actual input against a baseline. We again select `[MASK]` as the baseline token, and create baselines based on the length of the original input. Gradients are computed along a linear path, from baseline to actual input, representing a transition from absence of features to the actual input. The gradients are accumulated at multiple steps along the path. The result is a vector for each token, representing a separate gradient value for each of a feature's dimension. We convert these vectors into a single score per token by applying $L2$ normalization (Ray Choudhury et al., 2022).

Applying each saliency method results in four scores per token, representing the individual token's impact on a specific class of the MATRES dataset. We aggregate these scores into a single value by summing [4]. over each score. The resulting score indicates the token's overall significance across all classes. Special tokens, introduced during fine-tuning (§ 4.3), must be carefully considered. For IG, the special tokens are included in the baseline inputs, to ensure the integrity of the input. For Occlusion, they are not perturbed, allowing to measure the impact of regular tokens on the representation of the special tokens, which in turn affects the model's predictions.

[4] Summing or averaging are common approaches for representing the influence of a token across classes (Molnar, 2022; Atanasova et al., 2020a). Both might overlook the importance of tokens that are particularly influential for a specific class.

	Alignment	
	IG	Occlusion
LUKE $_{\text{large}}$	0.19	0.20
LUKE $_{\text{base}}$	0.21	0.18
RoBERTa $_{\text{large}}$	0.11	0.21
RoBERTa $_{\text{base}}$	0.27	0.17
BERT $_{\text{large-uncased}}$	0.52	0.25
BERT $_{\text{base-uncased}}$	0.56	0.28

Table 6: Alignment score between correctly predicted portions of counterfactual instances and saliency methods for each model.



[CLS] in competitions against the clock , some athletes display an ability to seize control .
think of the clark - kent - to - superman routines that john el ##way and michael jordan
often pulled in the final seconds . but ira ##m leon stands on the side ##lines of his own
race against time . (...) medical science is advancing at a rate that doesn ' t pre ##cl ##ude
the development of a treatment , but it ' s not clear if it will come in time . " no one knows
what technology will be available in five years , " said allan friedman , duke university
hospital ne ##uro ##sur ##ge ##on in chief , who in 2011 removed as much of leon ' s
brain tumor as possible . (...) " but leon can still run . two years after his brain - cancer
diagnosis , he recently ran a sub - five - minute mile for the first time since high school .
what has startled the medical community even more is what leon did this month in
beaumont , texas . he [a1] won [/a1] the gus ##her marathon , finishing in 3 : 07 : 35 . that
was one second slower than his personal record in the 26 . 2 - mile event , [a2] set [/a2]
days before he underwent brain surgery in early 2011 . [SEP]

Figure 3: Visualization of saliency scores obtained by occlusion for one instance, performed on BERT-large. The model correctly predicts the temporal relation between ”won” and ”set”. More saturated tokens indicate higher saliency.

4.8 Explanation Alignment Score

Recalling § 3.2, explanation alignment happens when a model accurately predicts both counterfactual and original instances, using the right cues, as indicated by saliency scores. We calculate an alignment score from the 300 instances where both original and counterfactual predictions are accurate. The score reflects the proportion of instances where the positive partition of tokens has a statistically significant higher average saliency score than the negative partition, suggesting reliance on correct information[5]. We use a one-tailed independence T-test at a 0.05 p-value to assess statistical significance, testing the null hypothesis that positive tokens do not have higher average saliency scores than negative ones, as per Ray Choudhury et al. (2022).

5 Results & Analysis

5.1 Model Evaluation

Table 4 shows the performance of fine-tuned models on the MATRES dataset. Across all models, weighted F1-scores consistently exceed macro F1-scores, indicating challenges in predicting minority classes, such as VAGUE. LUKE and RoBERTa models exhibit similar performance metrics, with their larger variants showing marginal improvements. However, these improvements are limited. BERT models show similar trend in improved performance when scaled, but they underperform relative to other models. This suggests that the notion, that larger models might perform better for some use cases (Zhong et al., 2021; Desai and Durrett, 2020), only partially holds true for a temporal relation classification on the MATRES dataset.

5.2 Counterfactual Evaluation

Table 5 shows a comparison of F1-weighted average scores for the selected models on 300 original versus counterfactual instances. For all models, we observe a significant decrease in performance on counterfactual instances compared to the original instances, with an average performance drop of 20%. This indicates overall challenges in maintaining expected reasoning when the conditions change. Contrary to expectations, larger model variants show a bigger performance drop. This indicate that larger models are less likely to perform well on altered inputs than their smaller variants. Future work could consider relaxing the criteria that a model's prediction on a counterfactual scenario must perfectly align with the true class. Instead, by analyzing prediction probabilities, we might show that models appropriately adjust their probabilities in response to counterfactual changes. This is particularly valuable for classification with unbalanced distribution of labels (Molnar, 2022).

5.3 Explanation Alignment

Table 6 shows the explanation alignment score between correctly predicted counterfactual instances against the two selected saliency-based methods. We observe that IG and Occlusion do not agree on the alignment scores. This lack of agreement between the two methods is consistent with previous findings (Ray Choudhury et al., 2022; Atanasova

[5] For a single instance with a random partition of tokens, the positive and negative partitions should have similar saliency scores. For a dataset this translates to them being significantly different in ≈ 0% of cases.

et al., 2020a), and it must be addressed to draw appropriate conclusions.

The alignment scores with IG indicate that smaller models, when making correct predictions for both original and counterfactual cases, are more likely to rely on relevant information compared to larger models. In contrast, Occlusion shows no consistent trend across model sizes, with scores that do not favor either smaller or larger models.

Potential interpretations for this inconsistency have been suggested. One interpretation is that IG may struggle to compute accurate saliency scores due to the discrete nature of text data (Harbecke and Alt, 2020), as the intermediate representations required do not align well with discrete word embeddings (Zhao et al., 2023), and therefore the computed gradients might not produce truthful saliency scores. Occlusion, potentially more stable, demonstrates no clear trend favoring model sizes. Another possible interpretation is that IG are in fact more faithful (Ray Choudhury et al., 2022). The trend shown by IG suggests that as the model's size increases, the features we define as important do not align with the model's strategies for correct predictions. Larger models, with their increased capacity, might be more likely to learn complex statistical patterns in the training data, including spurious ones. If the training data contain many such correlations, a larger model might be more prone to learn them and use them for predictions (Linzen, 2020). This could explain the higher accuracy of larger models compared to smaller ones (§ 5.1), but also indicates that larger models might depend on spurious patterns instead of relevant information (essentially, being right for the wrong reasons).

Overall, while the reasons behind inconsistencies remain unclear, the findings question the reliability of the selected saliency-based methods in evaluating model reasoning. Further work might include alternatives for computing saliency scores, such as surrogate models LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017).

6 Discussion

This study investigates selected discriminative models on a temporal relation classification task. While numerous benchmarks have been developed to evaluate models' temporal processing abilities, our experiments highlight limitations in these evaluations. Specifically, we adopted one commonly used benchmark dataset and found that models can achieve high accuracy without following the expected reasoning patterns. The framework used in this study offers a step toward improving evaluation methodologies by emphasizing whether models make correct predictions for the right reasons. It establishes clear success criteria for the task and highlights the role of validating "reasoning" to accurately assess model performance.

Post-hoc and local explainability methods are commonly used to determine if model decisions are justifiable from a human perspective, yet their reliability and utility is often questioned (Dasgupta et al., 2022; Saini and Prasad, 2022). Counterfactual explanations are considered as more truthful (Zhao et al., 2023), but require careful handling to prevent unreliable conclusions. Saliency scores, on the other hand, may not reflect the model's decision-making process. Different saliency methods can produce conflicting results, meaning that they inconsistently reflect the model's decision process (Jukić et al., 2023; Ding and Koehn, 2021b; Atanasova et al., 2020b). Moreover, the lack of a ground truth for saliency evaluation makes it challenging to evaluate whether they correctly approximate the model's processes (Molnar, 2022).

Having addressed the truthfulness of these methods, the question of their utility remains. For this study, we must conclude that the models follow some other strategy for correct prediction (rather than relying on the expected reasoning). Explainability methods should aim to make a model's decisions understandable to humans. However, this is challenging when a model's reasoning processes do not align with human reasoning (González et al., 2021). Identifying alternative reasoning strategies or shortcuts through these explanations is challenging because they are not necessarily human interpretable (see Figure 3[6]), raising questions about the practical value of these methods, as they only provide a partial interpretable view of a model's processes, and fail to provide actionable insights.

[6] Similar work (Ray Choudhury et al., 2022; Du et al., 2021) report both negative and positive impacts on saliency scores, which we consider as positive contributions regardless of probability direction.

7 Extending to Generative Models

Extending the experiments to adapt modern generative models, such as LLaMA (Touvron et al., 2023), GPT (Yenduri et al., 2023), and OLMO (Groeneveld et al., 2024), presented challenges, particularly in interpreting saliency scores.

Zhao et al. (2023) provides a taxonomy of explainability methods for transformer-based language models, categorizing them based on training paradigms (e.g., fine-tuning and prompting), which influence their goals and effectiveness. Generative models, primarily prompt-based, leverage their extensive scale and learned prompts for task execution. These complex processing strategies (Wei et al., 2023) make it difficult to isolate specific components of the model responsible for particular decisions. Localized and example-based explainability methods become less meaningful (Zhao et al., 2023). Moreover, differences in training objectives (e.g. autoregressive versus masked language), make it challenging to apply explainability methods that work reliably across all model types. Trustworthiness of explanations is both task and model-dependent (Bastings et al., 2022). Variations in how models process and prioritize input can result in inconsistencies in the effectiveness of these methods. This variability underscores that no single explanation method can be universally treated as a standard across all contexts. Consequently, conducting meaningful comparisons between different architectures becomes challenging, as the results may be unreliable or even misleading. Further research is needed to validate the robustness of such comparative analyses.

In contrast, counterfactual explanations provide a promising approach for evaluating generative models. Assessments centered on counterfactual instances could help determine whether these models maintain consistent reasoning when confronted with alternative scenarios. We leave the adaption of the presented counterfactual explanations (§ 4.6) to generative models to future work.

Of particular relevance, Roccabruna et al. (2024) highlights the performance gap between generative and discriminative models in temporal relation classification tasks. Encoder-only models based on RoBERTa consistently outperform generative models like LLaMA. This performance gap is attributed to RoBERTa's ability to fully utilize input context via masked language modeling, in contrast to LLaMA's autoregressive objective, which tends to prioritize final tokens in the input sequence. This underscores the significance of discriminative models for TRC and reinforces the value of evaluating whether their decisions are based on valid and expected reasoning patterns.

8 Conclusion

Temporal annotations are used to mark all linguistic features that express temporal information in text. We evaluate selected discriminative models on a temporal relation classification task, examining whether they rely on these features for correct predictions. Experiments involve a combination of counterfactual explanations and saliency-based methods. High alignment between these two explanations indicates that a model is following a valid processing strategy. We find that this is not the case for the selected models, meaning that they might learn spurious correlations or shortcuts rather than relying on the defined linguistic features that form temporal meaning. We evaluate the limitations of this framework by examining the utility of the explainability methods used, together with challenges and potential directions for extending the framework to generative models.

Limitations

This study focuses on a single dataset and task, which limits the generalizability of its findings. Future work could expand the scope by exploring additional benchmark datasets and tasks to assess the broader applicability of the proposed framework. Generating and testing a larger number of counterfactual and original instance pairs would also provide a more robust evaluation.

Our approach to saliency scores may additional attention. The current methodology does not account for the potential negative impact of individual tokens on predictions, and it aggregates all scores without identifying specific tokens that are particularly influential for a given class.

Acknowledgment

Many thanks to Anna Rogers for her invaluable guidance and support.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020b. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. “will you find these shortcuts?” a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding?

C. Callender. 2011. *The Oxford Handbook of Philosophy of Time*.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions.

Kiel Christianson. 2016. When language comprehension goes wrong for the right reasons: Good enough, underspecified, or shallow language processing. *Quarterly journal of experimental psychology (2006)*, 69:1–29.

Divyanshu Daiya. 2020. Combining temporal event relations and pre-trained language models for text summarization. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 641–646.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. 2022. Framework for evaluating faithfulness of local explanations.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding and Philipp Koehn. 2021a. Evaluating saliency methods for neural language models.

Shuoyang Ding and Philipp Koehn. 2021b. Evaluating saliency methods for neural language models.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu model.

Anoushka Gade and Jorjeta Jetcheva. 2024. It's about time: Incorporating temporality in retrieval augmented language models.

Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Harbecke and Christoph Alt. 2020. Considering likelihood in NLP classification explanations with occlusion and language modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 111–117, Online. Association for Computational Linguistics.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774, Singapore. Association for Computational Linguistics.

Josip Jukić, Martin Tutek, and Jan Šnajder. 2023. Easy to decide, hard to agree: Reducing disagreements between saliency methods.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316, Singapore. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

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

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia. Association for Computational Linguistics.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, Online. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.

Christoph Molnar. 2022. *Interpretable Machine Learning*. LeanPub.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023. Are large language models temporally grounded?

Adel Rahimi and Shaurya Jain. 2022. Testing the effectiveness of saliency-based explainability in nlp using randomized survey-based experiments. *ArXiv*, abs/2211.15351.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models “understand” language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ”why should i trust you?”: Explaining the predictions of any classifier.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification?

Anna Rogers, Marzena Karpinska, Ankita Gupta, Vladislav Lialin, Gregory Smelkov, and Anna Rumshisky. 2022. Narrativetime: Dense temporal annotation on a timeline.

Aditya Saini and Ranjitha Prasad. 2022. Select wisely and explain: Active learning and probabilistic local post-hoc explainability.

Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22. ACM.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Y. Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. 2023. Language models can improve event prediction by few-shot abductive reasoning.

Nina Spreitzer, Hinda Haned, and Ilse van der Linden. 2022. Evaluating the practicality of counterfactual explanations. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2019. Assessing the benchmarking capacity of machine reading comprehension datasets.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Xia Sitt Fankhauser Chicas-Mosier Monteith Tai, Bentley. 2024. An examination of the use of large language models to aid analysis of textual data.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. Towards benchmarking and improving the temporal reasoning capability of large language models.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023c. Towards benchmarking and improving the temporal reasoning capability of large language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention.

Jing Yang, Yu Zhao, Linyao Yang, Xiao Wang, Long Chen, and Fei-Yue Wang. 2024. Temprompt: Multi-task prompt learning for temporal relation extraction in rag-based crowdsourcing systems.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Guy Yanko, Shahaf Pariente, and Kfir Bar. 2023. Temporal relation classification in Hebrew. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 261–267, Nusa Dua, Bali. Association for Computational Linguistics.

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. 2023. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390, Seattle, United States. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey.

Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. 2021. Are larger pretrained language models uniformly better? comparing performance at the instance level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. RSGT: Relational structure guided temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2001–2010, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Appendix A

BERT (bert-base-uncased, bert-large-uncased), RoBERTa (FacebookAI/roberta-base, FacebookAI/roberta-large), LUKE (studio-ousia/luke-base, studio-ousia/luke-large) are sourced from the Hugging Face Transformers library. Each encoder model is fine-tuned for the task of temporal relation classification using the architectural and tokenisation strategies presented by Yanko et al. (2023) and Baldini Soares et al. (2019). All models are fine-tuned for the duration of 10 epochs with a batch-size of 8, using AdamW optimizer. The learning rate was kept at 1e-05.

Appendix B

	Relaxed F1	Relaxed F1	EM
	M/avg	W/avg	
LUKE $_{\text{large}}$	0.61	0.81	0.80
LUKE $_{\text{base}}$	0.61	0.80	0.78
RoBERTa $_{\text{large}}$	0.67	0.82	0.81
RoBERTa $_{\text{base}}$	0.65	0.81	0.79
BERT $_{\text{large-uncased}}$	0.66	0.81	0.78
BERT $_{\text{base-uncased}}$	0.63	0.79	0.77

Table: Performance evaluation on MATRES (Ning et al., 2018) dataset, using the ”relaxed” F1 metric proposed by Yanko et al. (2023).

VAGUE class was initially introduced in MATRES dataset to account for disagreements that arise during the annotation process (Ning et al., 2018). Yanko et al. (2023) introduces a ”related F1” metric to address the complexities associated with the class. This evaluation metric excludes errors where non-VAGUE predictions are made on VAGUE samples, based on the argument that VAGUE inherently encompasses both temporal directions (BEFORE and AFTER). Errors in this class are considered less critical and can be partially disregarded. Similarly, Roccabruna et al. (2024) take this notion further by completely excluding the VAGUE class from analysis, arguing that it does not represent a true temporal relation. We chose to keep the VAGUE class due to its potential value in generating counterfactual explanations. The class can serve as a middle ground that can be modified into more definitive temporal relations (BEFORE, AFTER or EQUAL) or created by introducing ambiguities into otherwise clear relationships.

Appendix C

This section provides a detailed overview of the methods used to generate counterfactual explanations, including how alterations were identified and implemented to ensure semantic correctness. Four types of possible and semantically correct alterations were employed to generate counterfactual explanations:

1. We consider simple temporal relationships those that contain explicit temporal conjunctions (e.g. ”before”, ”after” and ”while”). For simple temporal relationships, revering the temporal conjunction and/or changing verb tenses were sufficient as semantically correct alterations. This strategy most often resulted in reversing BEFORE and AFTER relationships.

2. For instances where a direct reversal of temporal conjunction or verb tense change was not possible, temporal conjunctions or adverbs (e.g. ”subsequently”, ”already”) and temporal expressions (e.g. ”months”, ”years”) were added or removed. This strategy often resulted in altering BEFORE or AFTER relationships to an EQUAL relationship, or vice-versa.

3. We consider more complex relationships those that include conditional or causal relationships between the two events. Focus was put in not altering the nature of such relationships. For these cases, reversing the temporal relationship involved reversing the cause with the effect or vice-versa.

4. For actions described in separate sentences, reordering the sentences was considered as a valid semantic alteration. This alteration is possible and particularly relevant for the dataset at hand, which is based on news snippets. For the news domain, the order of mention often dictates the sequence of events. This strategy often resulted in altering to or from a VAGUE relationship. Reordering sentences within the text, by placing them closer or further apart, either increased or decreased the contextual dependency between a pair of actions.

	Common Features	**Examples**
Temporal Expressions: Tokens that specify points in time	Absolute expressions, such as *December 2025, at 5PM* Relative expressions, such as *week, Mondays, annually*	She started a new job on September 1st, after moving to the city. If it rains tomorrow, the picnic will be postponed until Sunday.
Temporal Prepositions and Adverbs: Tokens used to connect actions or events to specific times.	Prepositions such as *at, on, in, during, for, over, by* Adverbials such as *again, late, now, then eventually, previously, recently*	She started a new job on September 1st, after moving to the city. Recently, he has taken up running before breakfast at 8AM.
Temporal Conjunctions: Tokens used to related events to each one another.	Conjunctions such as *before, after, while, until, since* *when, as soon as, as long as*	She started a new job on September 1st, just after moving to the city. Recently, he has taken up running before breakfast every morning.
Subordinate Conjunction: Tokens used to express conditional or causal relationship between events or actions.	References to causality such as *because, therefore, as* References to conditions such as *if, unless, then, so*	Because you didn’t reply in time, I only bought tickets for two. If it rains tomorrow, then the picnic will be postponed until Sunday at noon.

Appendix D: Examples of features that express temporal information. The table is designed to demonstrate how relevant and important tokens are identified and retrieved in accordance with the annotation guidelines. Color coding follows the annotation guidelines from TimeML (Mani et al., 2006): orange is used for signal tokens (`SIGNAL`), providing cues for how events and temporal expressions are related to each other; blue is used for specific time expressions (`TIMEX3`).