

WISTERIA: Weak Implicit Signal-based Temporal Relation Extraction with Attention

Duy Dao Do, Anaïs Halftermeyer, Thi-Bich-Hanh Dao
University of Orléans, INSA Centre Val de Loire, LIFO EA 4022, France
{duy-dao.do, anais.halftermeyer, thi-bich-hanh.dao}@univ-orleans.fr

Abstract

Temporal Relation Extraction (TRE) requires identifying how two events or temporal expressions are related in time. Existing attention-based models often highlight globally salient tokens but overlook the pair-specific cues that actually determine the temporal relation. We propose WISTERIA (Weak Implicit Signal-based Temporal Relation Extraction with Attention), a framework that examines whether the top- K attention components conditioned on each event pair truly encode interpretable evidence for temporal classification. Unlike prior works assuming explicit markers such as *before*, *after*, or *when*, WISTERIA considers signals as any lexical, syntactic, or morphological element implicitly expressing temporal order. By combining multi-head attention with pair-conditioned top- K pooling, the model isolates the most informative contextual tokens for each pair. We conduct extensive experiments on TimeBank-Dense, MATRES, TDDMan, and TDDAuto, including linguistic analyses of top- K tokens. Results show that WISTERIA achieves competitive accuracy and reveals pair-level rationales aligned with temporal linguistic cues, offering a localized and interpretable view of temporal reasoning.

Keywords: Temporal Relation Extraction, attention-based, Cross Attention, Multi-Head Attention, top- K Pooling, BERT

1. Introduction

Temporal Relation Extraction (TRE) is a key task in natural language processing (NLP) that identifies and classifies temporal relationships between events and temporal expressions in text. These relationships reveal temporal dynamics such as event order, duration, and causality. Understanding them aids information retrieval, event prediction, and knowledge graph construction. A main challenge in TRE is identifying which contextual elements encode the temporal relationship between two entities, as this directly affects accuracy and interpretability.

Traditional frameworks such as TimeML (Pustejovsky et al., 2006) introduce *signal words* (e.g., *before*, *after*, *when*) to mark explicit temporal links. Yet, such markers appear sparsely in real-world text. Annotators frequently infer relations using implicit linguistic cues, often through syntactic or common-sense reasoning. For example, in the TimeBank-Dense corpus (Cassidy et al., 2014), the pairs (*takeover*, *news*) and (*spent*, *thought*) are connected by explicit signals (*before*, *when*), while (*spent*, *sold*) or (*said*, *sold*) require implicit reasoning from context or syntax. This observation motivates our hypothesis that many temporal relations are carried by *implicit linguistic signals* beyond explicit markers.

Recent works in TRE leverage pretrained language models (PLMs) (Lin et al., 2019; Man et al., 2022), graph reasoning (Zhang et al., 2022; Mathur et al., 2021; Zhao et al., 2021), logic-based inference (Zhou et al., 2020; Huang et al., 2023), and

Not that long ago, **before** the Chinese **takeover**, the **news** about real estate here was that the sky was the limit the highest prices in the world. So **when** Wong Kwan **spent** seventy million dollars for this house, he **thought** it was a great deal. He **sold** the property to five buyers and **said** he'd double his money.

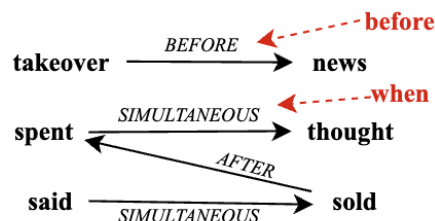


Figure 1: An example from TimeBank-Dense (Cassidy et al., 2014), based on TimeBank 1.2 (Pustejovsky et al., 2006). Explicit signals link *BEFORE*(*takeover*, *news*) and *SIMULTANEOUS*(*spent*, *thought*), while *spent-sold* depends on implicit contextual cues.

external knowledge integration (Ning et al., 2019; Han et al., 2020). Although effective, these methods are computationally costly and provide limited transparency into how contextual cues drive predictions. Few studies explicitly examine whether attention mechanisms can reveal the contextual rationale behind a predicted relation for each entity pair.

In this work, we introduce **WISTERIA**, a lightweight and interpretable framework that shifts

attention from global salience modeling to **relation-conditioned evidence modeling**. Instead of computing attention uniformly across the sentence, WISTERIA explicitly conditions attention on each entity pair and retrains only the top- K contextual tokens most influential.

Importantly, our goal is not to claim that attention constitutes a complete explanation of model decisions. Rather, we investigate whether relation-conditioned attention distributions align with systematically meaningful temporal features. To this end, we introduce a structured linguistic interpretability framework that analyzes the part-of-speech (POS), dependency (Dep), and morphological (Morph) properties of the selected top- K tokens. This analysis allows us to examine whether the extracted evidence corresponds to known temporal phenomena such as tense marking, aspectual morphology, and subordinating constructions.

We evaluate WISTERIA on four benchmark datasets: **TimeBank-Dense** (Cassidy et al., 2014), **MATRES** (Ning et al., 2019), and the discourse-level corpora **TDDMan** and **TDDAuto** (Naik et al., 2019). Experimental results demonstrate competitive performance, particularly on sentence-level datasets, while providing consistent linguistic alignment between attention-selected evidence and temporal cues.

Our contributions are threefold:

- We introduce a **relation-conditioned top- K attention mechanism** that models attention as entity-pair-specific evidence selection rather than global token salience.
- We propose a **linguistic interpretability protocol** that systematically analyzes POS, dependency, and morphological distributions of attention-selected tokens to evaluate temporal signal alignment.
- We demonstrate that **WISTERIA** achieves competitive performance across four TRE benchmarks while maintaining computational efficiency and offering structured, linguistically grounded interpretability.

2. Related Work

Temporal Relation Extraction has evolved through multiple paradigms, including pretrained language models (PLMs), graph-based reasoning, logic-based inference, and knowledge-enhanced approaches (Lin et al., 2019; Zhang et al., 2022; Mathur et al., 2021; Ning et al., 2019; Huang et al., 2023). While these methods achieve strong performance, they often require complex architectures or structured inference pipelines, limiting interpretability and computational efficiency.

Attention in TRE. Attention mechanisms (Bahdanau et al., 2014; Vaswani et al., 2017) are widely used in Transformer-based TRE models. However, conventional attention typically distributes weights according to sentence-level salience, which may dilute pair-specific signals. Selective context modeling and graph attention (Man et al., 2022; Zhang et al., 2022) improve performance but do not explicitly analyze whether attention aligns with linguistic temporal cues.

Top- K Attention and Evidence Selection.

Sparse and top- K attention mechanisms have been proposed for efficiency and interpretability in NLP (Correia et al., 2019; Child et al., 2019). In relation extraction, top- K evidence selection has been applied at the document level (Ma et al., 2023; Yuan et al., 2025). However, these approaches operate globally and are not conditioned on specific entity pairs.

Our Positioning. WISTERIA differs by explicitly conditioning attention on entity-pair semantics and analyzing the linguistic properties of the selected evidence. By combining relation-conditioned top- K attention with structured linguistic analysis, our framework bridges selective evidence modeling and interpretable temporal reasoning.

3. Preliminaries

3.1. Background

Attention mechanisms (Bahdanau et al., 2014; Vaswani et al., 2017) have become a cornerstone of modern NLP by enabling models to dynamically focus on contextually relevant parts of the input. This mechanism lies at the heart of the Transformer architecture (Vaswani et al., 2017), which forms the foundation for most state-of-the-art pretrained language models. Transformer-based models such as BERT and RoBERTa (Devlin et al., 2018; Liu et al., 2019) leverage attention to generate rich contextual embeddings that capture deep semantic and syntactic dependencies between tokens. However, these models remain constrained by fixed input length (typically 512 tokens) and limited interpretability-attention weights do not always correspond to meaningful linguistic evidence (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019).

To improve efficiency and interpretability, several variants such as multi-head and top- K attention (Correia et al., 2019; Child et al., 2019) have been proposed. While top- K attention sparsifies focus by selecting the most relevant tokens, it generally operates globally without conditioning on specific relational pairs. Motivated by this limitation, we

introduce a more selective and interpretable mechanism, pair-conditioned top- K attention, built upon the BERT base architecture to enhance contextual focus at the entity-pair level.

3.2. Attention Pooling

To obtain a single contextual embedding from token representations $X \in \mathbb{R}^{n \times d}$, we use a simple attention-based pooling mechanism-motivated by prior uses of attention for sequence summarization (Zhuoran et al., 2021; Safari et al., 2020) - defined as

$$att_pooling(X) = Att(m_X, X, X), \quad (1)$$

where m_X is the mean of X used as the query. This instantiation highlights tokens most relevant to the context with complexity $O(nd)$. We then extend this idea to pair-conditioned top- K pooling, which selectively aggregates tokens most informative for each entity pair to improve interpretability and efficiency.

4. Methodology

Our proposed framework combines contextual representation learning with selective attention to improve temporal relation extraction. Figure 2 illustrates the overall architecture, consisting of four main components: context construction, multi-level embedding extraction, pair-conditioned top- K attention, and relation classification.

4.1. Context Construction

Given two entities e_1 and e_2 in a document, we dynamically extract a context window C using a predefined window size ws . When the entities are close, a single centered window is used; when they are far apart, two subwindows centered on each entity are concatenated, ensuring both entities appear within C . Each entity span is marked with four special tokens: $[E1], [/E1], [E2], [/E2]$, following prior temporal relation works (Ning et al., 2019; Lin et al., 2019; Dligach et al., 2017; Knez and Žitnik, 2024).

The marked text is tokenized by both a Transformer-based tokenizer (BERT or RoBERTa) and the *spaCy* tokenizer (Honnibal and Johnson, 2015). Subword-to-token alignment is performed using the *spaCy Alignment* module, enabling the construction of (i) **entity masks** for e_1 and e_2 , and (ii) **word-level masks** linking *spaCy* tokens to BERT subwords. These masks guide the subsequent attention-based pooling operations.

4.2. Multi-level Embedding Extraction

Let $C = [w_1, w_2, \dots, w_n]$ denote the token sequence. We first obtain subword embeddings

$X \in \mathbb{R}^{n \times d}$ from a pretrained BERT encoder. To derive word-level embeddings, we apply an attention-based pooling mechanism (Zhuoran et al., 2021; Safari et al., 2020) over the subwords belonging to each *spaCy* token:

$$h_i^w = Att(m_i, X, X), \quad (2)$$

where m_i is the mean vector of subwords of word i used as the query. Similarly, entity-level embeddings h_{e_1} and h_{e_2} are obtained by aggregating the word embeddings within each marked entity span using the same formulation.

The resulting word embeddings are further refined through a lightweight Transformer encoder with few layers of multi-head self-attention and learnable positional encoding (Shaw et al., 2018), producing contextualized representations $H^c = [h_1^c, \dots, h_M^c]$.

4.3. Pair-conditioned Top- K Attention

Given the entity-level representations h_{e_1} and h_{e_2} , we first construct a relation-aware pair representation by concatenation followed by linear projection:

$$h_{pair} = [h_{e_1} \oplus h_{e_2}]W_p, \quad (3)$$

where W_p projects the concatenated entity embedding into the contextual space. This projection aligns the pair representation with contextual word representations, enabling relation-specific attention computation.

Relation-conditioned cross-attention. Unlike conventional attention mechanisms that distribute weights based on sentence-level salience, our approach conditions attention distributions on entity-pair semantics. Specifically, we instantiate cross-attention using three query representations - e_1 , e_2 , and the projected pair embedding h_{pair} - following the multi-head attention formulation of Vaswani et al. (2017). For each query $q \in \{e_1, e_2, pair\}$, attention scores are computed over contextual representations H_c . Because each query encodes entity- or pair-specific semantics, the resulting attention distributions are relation-dependent rather than sentence-global, thereby functioning as pair-aware evidence selectors rather than generic token importance estimators.

Top- K evidence selection. To promote selectivity, we retain only the top- K contextual tokens with the highest attention weights for each query, following sparse attention principles (Correia et al., 2019; Child et al., 2019). Let H_q^k denote the selected tokens for query q , which are aggregated as:

$$h_q^k = Att(h'_q, H_q^k, H_q^k), \quad (4)$$

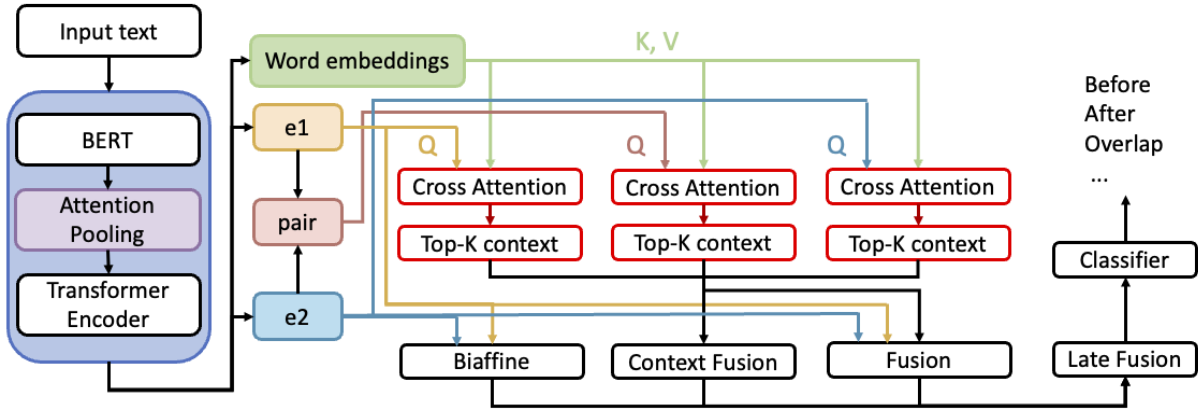


Figure 2: Architecture of WISTERIA. BERT and a transformer encoder generate contextualized representations, while pair-conditioned top- K cross-attention extracts key context for each entity pair. The biaffine, context, and pair-fusion outputs are integrated via late fusion for temporal relation classification.

where h'_q is the projected query representation.

Because selection is conditioned on entity-pair semantics, different pairs within the same sentence may focus on distinct contextual evidence. This contrasts with global top- K attention, where token selection is relation-independent, and enables pair-specific evidence modeling.

Label-aware gating. To enhance relational discriminability, we introduce learnable label embeddings (Ma et al., 2016; Rios and Kavuluru, 2018) as semantic prototypes for temporal relation classes. Instead of serving solely as output targets, these embeddings are incorporated into the representation space and fused with contextual features through a gating mechanism. This allows the model to modulate pair representations according to class-level semantics. The label embeddings act as soft guidance rather than hard constraints, complementing relation-conditioned evidence selection while preserving architectural simplicity.

4.4. Relation Classification

We employ three complementary prediction heads: (1) a biaffine scorer (Dozat and Manning, 2017) modeling direct interactions between e_1 and e_2 ; (3) a context head focusing solely on top- K contextual signals; and (2) a fusion head combining entity and contextual features. Their logits are integrated using a learnable late-fusion vector:

$$\hat{y} = \text{softmax} \left(\sum_{i=1}^3 w_i \cdot \text{logits}_i \right), \quad (5)$$

and the entire model is optimized using the cross-entropy objective.

Overall, our architecture achieves computational complexity $O(L(n^2d + d^2))$, comparable to Transformer-based encoders, while yielding more selective and interpretable attention distributions.

5. Experiments

The experiments aim to answer the following research questions:

- **Q1:** How effective is our model in capturing temporal relations compared to standard BERT-based baselines?
- **Q2:** To what extent is our model interpretable in revealing how contextual signals contribute to temporal reasoning?

5.1. Datasets & Evaluation

We evaluate our model on four benchmark datasets for temporal relation extraction. **TimeBank-Dense (TBD)** (Cassidy et al., 2014) densifies the original TimeBank corpus, adding more temporal links than human annotators would naturally provide. **MATRES** (Ning et al., 2019) simplifies TBD by reducing relation classes and removing weakly grounded links (not on the same "temporal axis"), and adding it to the **AQUAINT** (Graff, 2002) corpus, while **TDDMan** and **TDDAuto** (Naik et al., 2019) are new versions of TBD, introducing inter-sentential relations for discourse-level reasoning, respectively in a manual way, and automatically.

Table 1 summarizes the data distribution and label sets for all four datasets.

For all datasets, we report the standard Precision, Recall, and Micro-average F1 scores to

Dataset	Train	Validation	Test	Labels
TDDMan	4000	650	1500	a, b, s, i, ii
TDDAuto	32609	1435	4258	a, b, s, i, ii
MATRES*	10404	1836	817	e, a, b, v
TBD	4032	629	1427	a, b, s, i, ii, v

Table 1: Train/Validation/Test data distribution for TDDMan, TDDAuto, MATRES, and TimeBank-Dense. Label abbreviations: a = *After*, b = *Before*, s = *Simultaneous*, i = *Includes*, ii = *Is_included*, v = *Vague*, e = *Equal*. *(Ning et al., 2019) use TimeBank and Aquaint for training, Platinum for testing, and 20% of the training data as validation.)

ensure consistent comparison across models. Following common practice in temporal relation extraction, for TimeBank-Dense and MATRES we exclude the Vague label from evaluation, as done in previous studies (Huang et al., 2023; Mathur et al., 2021; Zhang et al., 2022; Xu et al., 2022; Yao et al., 2024).

5.2. Baseline Models

We compare our proposed model (**WISTERIA**) with a comprehensive set of strong baselines from recent work on temporal relation extraction across four benchmark datasets. The compared methods include: (1) **BiLSTM** (Cheng and Miyao, 2017), an LSTM-based architecture with discourse-level context; (2) fine-tuned **BERT-based** models as implemented in Ballesteros et al. (2020); (3) **Unified Framework** (Huang et al., 2023), combining pretrained language models with logical or structural constraints; (4) graph-based methods such as **TIMERS** (Mathur et al., 2021); (5) graph-distillation and contrastive learning approaches including **MuLCo** (Yao et al., 2024); (6) the **DTRE** model (Wang et al., 2022) for discourse-level reasoning; and (7) **CPTRE** (Yuan et al., 2024), a contrastive pretraining model for document-level temporal relation extraction.

Our model differs from these baselines by incorporating a pair-conditioned top- K cross-attention mechanism, which selectively focuses on the most informative contextual signals for each entity pair, enhancing both performance and interpretability.

5.3. Experimental Settings

We fine-tune the **BERT-base-uncased**¹ model (Devlin et al., 2018) with an additional single Transformer encoder layer to derive contextualized embeddings. The model is implemented in PyTorch using the fused AdamW optimizer (Loshchilov and Hutter, 2017). All experiments are conducted on a

¹<https://huggingface.co/google-bert/bert-base-uncased>

single NVIDIA Quadro GV100 GPU. Table 2 summarizes the key hyperparameters used in our experiments.

Setting	Value
Pretrained model	BERT-base-uncased
Encoder layer	1 Transformer layer
Optimizer	AdamW (fused)
Dropout rate	0.5
Learning rate	1×10^{-5}
Batch size	128
Epochs	30
Attention heads	8
Top- K range	[1-20]
FFN hidden size	512
Activation	GELU
Hardware	1 × Quadro GV100 GPU

Table 2: Hyperparameter configuration for training WISTERIA.

5.4. Results

Table 3 presents the F1-scores across four benchmark datasets. WISTERIA archives strong performance on sentence-level datasets (TBD and MATRES) and competitive results on discourse-level datasets (TDDAuto and TDDMan). Compared with early neural architectures such as BiLSTM, the model demonstrates substantial gains, confirming the benefit of contextualized representations combined with relation-conditioned evidence selection.

Against standard BERT-based baselines, WISTERIA yields notable improvements (e.g., +20.9 F1 on TBD) despite relying on a lightweight architecture with only one additional Transformer layer. When compared with graph-based or constraint-driven systems (TIMERS, DTRE, MuLCo, CPTRE), our model achieves comparable performance while maintaining architectural simplicity. This suggests that relation-conditioned attention provides an efficient mechanism for capturing temporal cues without requiring heavy structural modules.

Across datasets, WISTERIA performs best on TBD and MATRES, which primarily contain short-range, sentence-level relations. On document-level corpora, performance remains competitive but trails graph-based models that explicitly propagate long-distance dependencies. This distinction highlights the modeling scope of WISTERIA: it specializes in localized relational evidence selection rather than global constraint enforcement.

Figure 3 illustrates the variation of F1 scores with different Top- K values across datasets. Despite variations in K across corpora, the stable F1 scores suggest notable differences in discourse nature. TBD and MATRES form a coherent group-unsurprising since MATRES includes TBD-as both involve short-distance relations, whereas TDD

Model	PLM	TBD	MATRES	TDDAuto	TDDMan
BiLSTM (Cheng and Miyao, 2017)	word2vec	0.484	0.595	0.518	0.243
BERT-based (Ballesteros et al., 2020)	BERT-base	0.622	0.772	0.623	0.375
Unified-Framework (Huang et al., 2023)	RoBERTa-Large	0.681	0.826	–	–
TIMERS (Mathur et al., 2021)	BERT-Large	0.678	0.823	0.711	0.455
DTRE (Wang et al., 2022)	BERT-base	0.692	–	0.702	0.500
MuLCo (Yao et al., 2024)	BERT-base	0.814	0.857	0.662	0.473
CPTRE (Yuan et al., 2024)	BERT-base	0.714	0.842	0.807	0.568
WISTERIA (Ours)	BERT-base	0.831	0.843	0.709	0.4973

Table 3: Comparison of F1-scores across four benchmark datasets: TimeBank-Dense (TBD), MATRES, TDDAuto, and TDDMan. The best performance for each dataset is highlighted in bold.

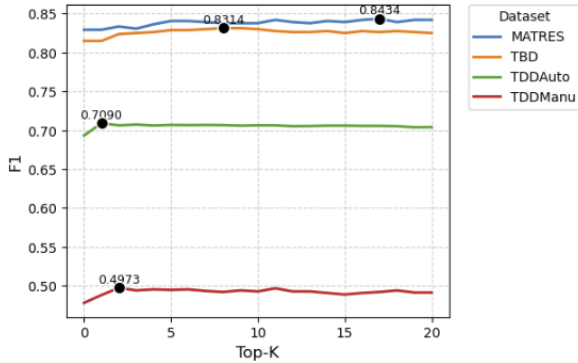


Figure 3: Effect of Top- K values on F1 performance across four datasets.

(Man and Auto) mainly exhibit long-distance ones. Figure 4 shows the distance between entity pairs across temporal datasets.

Overall, WISTERIA remains stable across a wide range of K , indicating that the pair-conditioned attention consistently selects informative contextual tokens without overfitting to specific thresholds. The optimal performance is achieved $K = 8$ for TBD (0.8314) and $K = 18$ for MATRES (0.8434), suggesting that moderate context aggregation may be sufficient for capturing most temporal cues in local event pairs.

For document-level datasets (TDDAuto, TDDMan), performance shows minimal improvement beyond $K = [1; 2]$, reflecting that adding distant context contributes little additional signal and may even introduce noise due to cross-sentence sparsity. This pattern confirms that local temporal relations benefit more from focused attention, while global reasoning requires richer discourse-level modeling beyond token-level selection.

5.5. Top- K Analysis

We do not claim that attention weights alone constitute fully faithful or causally complete explanations of model decisions. Rather, we evaluate whether relation-conditioned top- K attention ex-

hibits systematic alignment with linguistically recognized temporal cues. Our analysis therefore focuses on distributional consistency, structural enrichment, and cross-dataset stability, rather than on token-level causal attribution. The distribution of POS, Dependency and Morphological features of TimeBank-Dense, MATRES, TDDAuto, TDDMan are shown in Appendix (10).

5.5.1. Linguistic Analysis on Local Datasets

For TimeBank-Dense and MATRES, which primarily contain sentence-level temporal relations, we examine the linguistic composition of the selected top- K tokens.

POS In both datasets, nouns (`NOUN`) and proper nouns (`PROPN`) dominate the attention space of e_1 and e_2 about 20% and 12–13% respectively—indicating that the model concentrates on event-denoting tokens and their argument structures. For the pair-conditioned representation, the attention distribution shifts toward verbs (`VERB`, ≈ 11 –12%) and adpositions (`ADP`, ≈ 11 –12%), showing that the model emphasizes functional and relational markers (e.g., "before", "in", "during") when forming a temporal link between two events (TBD, 2572/18936 tokens); (MATRES, 2151/17016 tokens).

This pattern is consistent across TBD and MATRES, confirming that pair-conditioned top- K attention effectively identifies lexical items carrying temporal meaning.

Dependency The dependency distributions reveal distinct syntactic focuses across datasets. In TBD, temporal entities mainly align with nominal and predicate–argument structures, dominated by `amod`, `compound`, `det`, `nsubj`, `pobj`, and `prep` (≈ 7 –11%), reflecting event mentions embedded in noun phrases or governed by prepositions. In contrast, MATRES shows a similar but more clause-oriented pattern, with `compound`, `det`, `nsubj`, `pobj`, `prep`, and notably `punct` as top dependencies (≈ 8 –12%). The prominence of `punct` sug-

Dataset	Avg. Character Distance	Avg. Token Distance	Min Distance	Max Distance
TBD	116.99	19.51	1	366
MATRES	128.94	24.01	1	416
TDDAuto	800.99	127.61	2	3349
TDDMan	913.61	146.65	61	3074

Table 4: Average, Minimum, and Maximum Character and Token Distance between Entity Pairs across Temporal Datasets

gests stronger reliance on clause boundaries and coordination signals. Across both datasets, the pair-conditioned view amplifies connective dependencies such as `mark`, `advcl`, and `prep`, which frequently introduce temporal clauses (e.g., *when*, *as*, *after*) and facilitate pair-level temporal reasoning.

Morphological Morphological evidence reveals that WISTERIA is sensitive to tense and aspectual cues. Tokens marked with `Tense=Past` or `Aspect=Perf` occur frequently across all entity types (around 20–22% in TBD and 17–18% in MATRES), indicating that the model leverages verbal morphology to infer temporal ordering, particularly relations such as *BEFORE* and *AFTER*. In the pair-conditioned context, the distribution of `Tense=Past` and `Tense=Pres` becomes more balanced, suggesting that the model compares events situated in different temporal frames rather than focusing on a single clause’s internal tense. Overall, these patterns confirm that the top- k tokens are not randomly attended but encode distinct morphosyntactic evidence relevant to temporal reasoning.

5.5.2. Linguistic Analysis on Global Datasets

We extend our linguistic analysis to the two document-level subsets of the TDDiscourse corpus: TDDAuto and TDDMan. These datasets contain long-range and cross-sentence temporal relations, where explicit lexical signals are scarcer and temporal reasoning depends more on discourse-level structures.

POS The POS distributions of TDDAuto and TDDMan reveal a stronger reliance on nominal categories and a relative reduction in overt relational markers. Across both datasets, nouns (`NOUN`) are the most frequent category ($\approx 19\text{--}21\%$), followed by proper nouns (`PROPN`, $\approx 12\text{--}17\%$). These proportions are notably higher than those of verbs (`VERB`, $\approx 10\text{--}12\%$) and adpositions (`ADP`, $\approx 10\text{--}11\%$), showing that the model grounds temporal reasoning primarily in event anchors and discourse entities rather than explicit lexical connectors. The pair-conditioned attention shows modest increases in

`VERB` and `ADP`, but overall the token distribution remains noun-dominant-consistent with the more implicit temporal cues found in document-level narratives.

Dependency Dependency patterns further confirm this shift toward discourse-level inference. For e_1 and e_2 , core dependencies such as `nsubj`, `obj`, and `amod` occur most frequently ($\approx 7\text{--}10\%$), indicating that attention centers on syntactic heads and event arguments. In the pair-conditioned representation, temporal relations are often mediated through `prep`, `advcl`, and `relcl` ($\approx 9\text{--}11\%$), suggesting that the model selectively captures structural links between clauses or sentences. Compared to TBD and MATRES, however, the frequency of overt subordinating markers (e.g., `mark`, `advcl`) is lower, implying that WISTERIA must rely on more implicit contextual dependencies to infer relations across discourse boundaries.

Morphological At the morphological level, both TDDAuto and TDDMan exhibit diverse tense-aspect patterns reflecting narrative variation. Tokens with `Aspect=Perf|Tense=Past` remain prominent ($\approx 17\text{--}19\%$), but less dominant than in local datasets, suggesting that the model draws less from explicit verb morphology and more from contextual coherence. Forms tagged with `VerbForm=Inf` ($\approx 4\text{--}5\%$) and `VerbForm=Fin` ($\approx 1\text{--}2\%$) are more frequent, consistent with the narrative reporting style typical of multi-sentence documents. Overall, morphological cues contribute less distinctly to temporal ordering here, emphasizing the need for higher-level relational reasoning.

5.5.3. Summary

Across all four datasets, the linguistic analyses reveal a clear continuum between local and global temporal reasoning. In sentence-level datasets such as TBD and MATRES, WISTERIA anchors each event (e_1, e_2) in noun phrases and relies on verbal and prepositional tokens (e.g., *when*, *as*, *after*) together with tense/aspect morphology to establish explicit temporal links. In contrast, in document-level datasets such as TDDAuto and TDDMan, the model encounters fewer overt tempo-

to find a diplomatic solution. And I hope that, whatever happens today, that our relationships with Russia will continue to be **productive** and constructive and strong, because that's very important to the future of our peoples. One contrary **view** of the issue **presented itself** to the president **as** he arrived in Philadelphia later in the day. Nevertheless, **the president** [E1] **said** [/E1] Washington would use force if diplomacy fails to [E2] **force** [/E2] Saddam Hussein to **back** down. The Russian foreign minister, meanwhile, sought to soften the harsh words of his military counterpart, **saying** on Friday that Russia now feels the US must hold off at least until UN secretary general Kofi Annan **visits** Baghdad in a last-ditch effort at diplomacy. Annan has no trip planned so **far**. Meanwhile,

Ground-truth: *Before* Prediction: *Before*

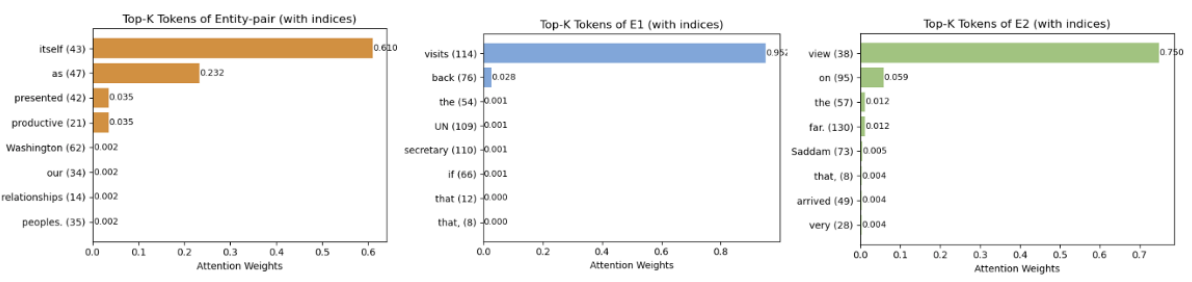


Figure 4: Example from the TimeBank-Dense test set illustrating pair-conditioned top- K attention. The model predicts *BEFORE* between *said* (E_1) and *force* (E_2). Pair-level attention highlights connective cues (e.g., *as*, *itself*), while entity-level attention anchors event-specific tokens (e.g., *visits*, *Saddam*). The complementary patterns indicate structured evidence selection for temporal inference.

ral connectives and instead integrates information from syntactic dependencies (*compound*, *amod*, *prep*) and discourse context that span multiple sentences. This indicates a shift from surface lexical cues to structural and contextual signals of temporal progression. Such transition from explicit to implicit temporal encoding explains why WISTERIA achieves higher F1 scores on local datasets, while maintaining slightly lower yet competitive performance (≈ 0.70) on global datasets-highlighting both its interpretability and adaptability across different reasoning scopes.

5.5.4. Interpretability

Figure 4 presents a representative example from TimeBank-Dense. Pair-level attention highlights connective and contextual framing tokens (e.g., *as*, *itself*, *productive*), whereas entity-level attention anchors event-specific roles. The complementary patterns indicate a functional distinction between event grounding and relational signaling.

We do not claim strict causal faithfulness. Rather, our analysis demonstrates structured and linguistically coherent evidence selection. Across datasets, attention distributions show (i) enrichment of temporal morphosyntactic features, (ii) stability across K values, and (iii) consistent specialization between entity-level and pair-level queries.

These findings suggest that relation-conditioned attention functions as a structured evidence filter rather than a diffuse salience mechanism. Establishing formal causal guarantees remains an open challenge, and integrating constraint-based reasoning over top- K evidence is a promising direction for future work.

6. Conclusion

We presented **WISTERIA**, a lightweight framework for temporal relation extraction that introduces relation-conditioned top- K attention for entity-pair-specific evidence selection. By conditioning attention on event-pair semantics, WISTERIA isolates informative contextual cues and achieves competitive performance across four benchmarks (TimeBank-Dense, MATRES, TDDAuto, and TDMan) while maintaining computational efficiency.

Beyond predictive accuracy, our linguistic analyses show that the selected top- K tokens consistently align with temporal morphosyntactic features, providing structured and transparent evidence selection. Rather than serving as a global salience mechanism, relation-conditioned attention functions as a pair-aware evidence filter.

Future work will integrate constraint-driven or rule-based reasoning on top of the extracted evidence to extend WISTERIA from pairwise classification toward globally consistent temporal reasoning.

To support reproducibility, we release the full implementation and top- K attention files².

7. Limitations

While WISTERIA advances relation-conditioned evidence selection and structured interpretability in temporal relation extraction, several limitations remain.

²<https://github.com/doduydao/WISTERIA>

Independent pairwise modeling. WISTERIA models each entity pair independently and does not incorporate explicit temporal constraints such as transitivity, antisymmetry, or global consistency enforcement. As a result, although pair-conditioned top- K attention effectively isolates locally informative contextual signals, it does not guarantee document-level temporal coherence. This limitation is particularly evident in discourse-level datasets (TDDAuto and TDDMan), where long-distance and multi-hop dependencies require structured reasoning beyond token-level evidence selection. Integrating constraint-driven or graph-based inference modules on top of the extracted evidence constitutes an important direction for future work.

Local evidence vs. global reasoning. The current top- K mechanism operates at the level of token selection conditioned on a specific entity pair. While this design enhances focus and interpretability, it does not explicitly propagate information across multiple pairs within a document. Consequently, WISTERIA excels at localized relational reasoning but is not designed as a full temporal reasoning engine. Future extensions may explore hybrid architectures that combine relation-conditioned attention with hierarchical, graph-based, or energy-based global reasoning frameworks.

Interpretability scope. The interpretability provided by WISTERIA is structural and distributional rather than strictly causal. Our analyses demonstrate systematic alignment between attention-selected tokens and linguistically recognized temporal cues; however, attention weights alone do not constitute provably faithful explanations in the causal sense. Establishing formal guarantees of explanation faithfulness remains a broader open challenge in neural interpretability research. We therefore position WISTERIA’s interpretability as a transparent evidence selection mechanism rather than a complete model of human-understandable reasoning.

Dependence on automatic linguistic annotation. The POS, dependency, and morphological analyses rely on automatic annotations, which may introduce noise. Although the consistency of observed patterns across datasets mitigates this concern, future work could investigate tighter integration between attention-selected evidence and rule-based linguistic validation.

Overall, WISTERIA should be viewed as a lightweight, relation-conditioned evidence extraction framework. Extending it with explicit temporal

constraints or symbolic reasoning components represents a promising pathway toward unified predictive and reasoning-based temporal understanding.

8. Ethics statement

This work did not raise ethical concerns during its development and is not expected to pose any in the future. From an ethical perspective, the contribution of this research lies in advancing toward more resource-efficient and interpretable systems. It is important to note that the knowledge extracted from these systems should not be conflated with human cognitive understanding of linguistic temporality. Rather, insights derived from neural models should be treated as investigative leads to be empirically tested against theories of human temporal processing. Furthermore, from a broader societal perspective, research that advances the interpretability of deep learning systems contributes to enhanced human oversight and control of these increasingly ubiquitous technologies in daily life.

9. References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. [Neural machine translation by jointly learning to align and translate](#). *arXiv preprint arXiv:1409.0473*.
- Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. [Severing the edge between before and after: Neural architectures for temporal ordering of events](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5412–5417, Online. Association for Computational Linguistics.
- Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. [An annotation framework for dense event ordering](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.
- Fei Cheng and Yusuke Miyao. 2017. [Classifying temporal relations by bidirectional lstm over dependency paths](#). pages 1–6.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. [Generating long sequences with sparse transformers](#). *arXiv preprint arXiv:1904.10509*.

- Gonalo M Correia, Vlad Niculae, and Andr  FT Martins. 2019. [Adaptively sparse transformers](#). *arXiv preprint arXiv:1909.00015*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. [Neural temporal relation extraction](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain. Association for Computational Linguistics.
- Timothy Dozat and Christopher D. Manning. 2017. [Deep biaffine attention for neural dependency parsing](#).
- David Graff. 2002. [The aquaint corpus of english news text](#).
- Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. [Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5717–5729, Online. Association for Computational Linguistics.
- Matthew Honnibal and Mark Johnson. 2015. [An improved non-monotonic transition system for dependency parsing](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. [More than classification: A unified framework for event temporal relation extraction](#).
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not explanation](#). In *North American Chapter of the Association for Computational Linguistics*.
- Timotej Knez and Slavko  itnik. 2024. [Multimodal learning for temporal relation extraction in clinical texts](#). *Journal of the American Medical Informatics Association*, 31(6):1380–1387.
- 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.
- 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 pre-training approach](#).
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). *arXiv preprint arXiv:1711.05101*.
- Youmi Ma, An Wang, and Naoaki Okazaki. 2023. [DREEM: Guiding attention with evidence for improving document-level relation extraction](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.
- Yukun Ma, Erik Cambria, and Sa Gao. 2016. Label embedding for zero-shot fine-grained named entity typing. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 171–180.
- Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. [Selecting optimal context sentences for event-event relation extraction](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11058–11066.
- Puneet Mathur, Rajiv Jain, Franck Deroncourt, 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.
- Aakanksha Naik, Luke Breitfeller, and Carolyn Penstein Ros . 2019. [Tddiscourse: A dataset for discourse-level temporal ordering of events](#). In *SIGDIAL Conferences*.
- 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.
- James Pustejovsky, Marc Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2006. Timebank 1.2. <https://catalog.ldc.upenn.edu/LDC2006T08>. LDC2006T08.
- Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2018, page 3132.
- Pooyan Safari, Miquel India, and Javier Hernando. 2020. [Self-attention encoding and pooling for speaker recognition](#). In *Interspeech*.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. [Self-attention with relative position representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in neural information processing systems*, pages 5998–6008.
- Liang Wang, Peifeng Li, and Sheng Xu. 2022. [DCT-centered temporal relation extraction](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2087–2097, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#).
- Xiaoliang Xu, Tong Gao, Yuxiang Wang, and Xinle Xuan. 2022. [Event temporal relation extraction with attention mechanism and graph neural network](#). *Tsinghua Science and Technology*, 27(1):79–90.
- Hao-Ren Yao, Luke Breifeller, Aakanksha Naik, Chunxiao Zhou, and Carolyn Rose. 2024. [Distilling multi-scale knowledge for event temporal relation extraction](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 2971–2980. ACM.
- Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2024. [Temporal relation extraction with contrastive prototypical sampling](#). *Knowledge-Based Systems*, 286:111410.
- Jiawei Yuan, Hongyong Leng, Yurong Qian, Jiaying Chen, Mengnan Ma, and Shuxiang Hou. 2025. [Evidence and axial attention guided document-level relation extraction](#). *Computer Speech & Language*, 90:101728.
- 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.
- Xinyu Zhao, Shih-Ting Lin, and Greg Durrett. 2021. [Effective distant supervision for temporal relation extraction](#). In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 195–203, Kyiv, Ukraine. Association for Computational Linguistics.
- Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2020. [Clinical temporal relation extraction with probabilistic soft logic regularization and global inference](#).
- Shen Zhuoran, Zhang Mingyuan, Zhao Haiyu, Yi Shuai, and Li Hongsheng. 2021. [Efficient attention: Attention with linear complexities](#). In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3530–3538.

10. Appendix

10.1. POS distribution

10.2. Dependency feature distribution

10.3. Morphological feature distribution

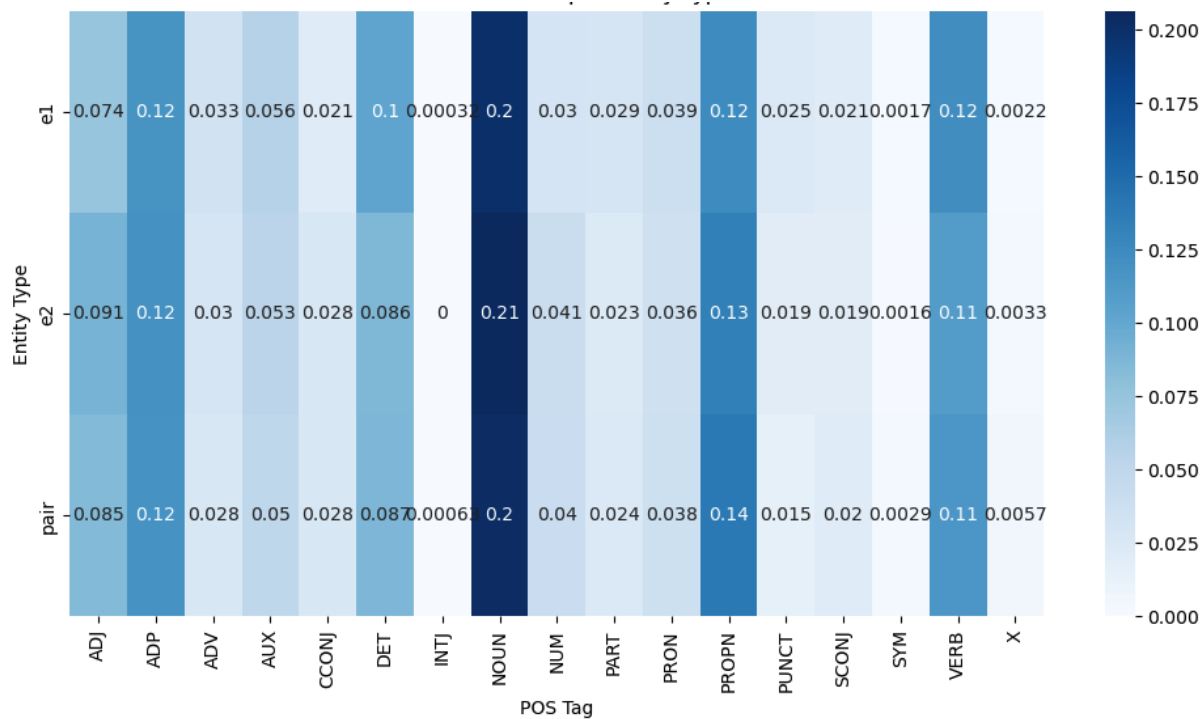


Figure 5: POS feature distribution of TimeBank-Dense

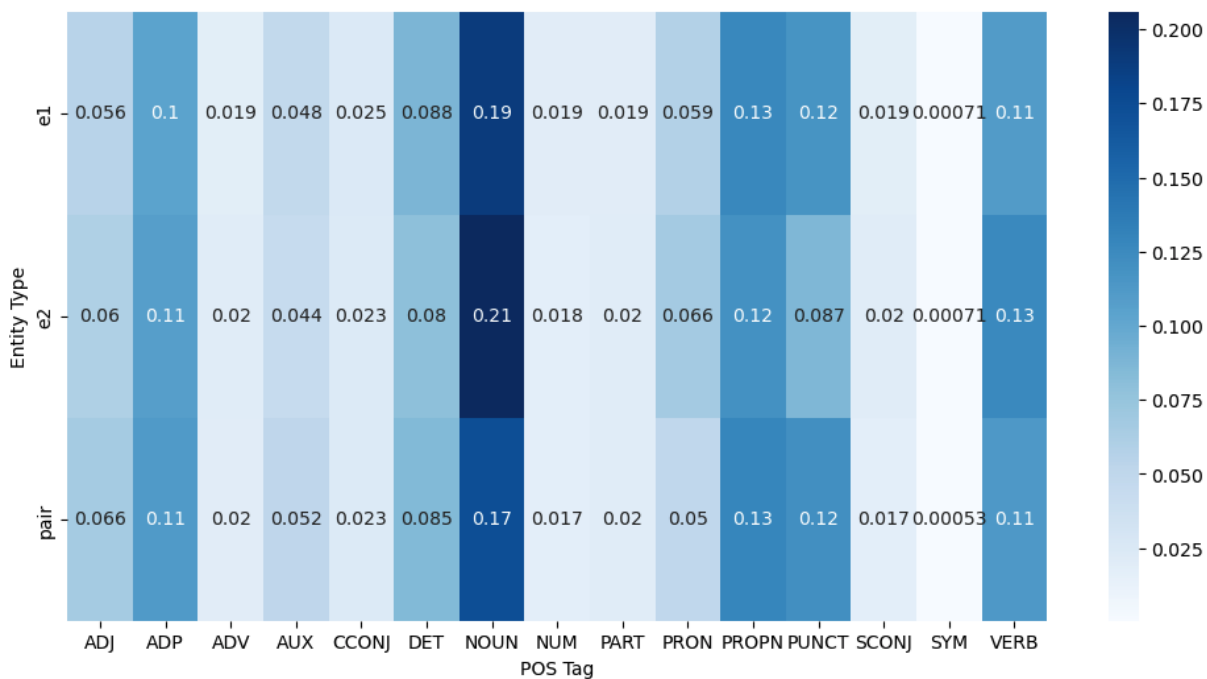


Figure 6: POS feature distribution of MATRES

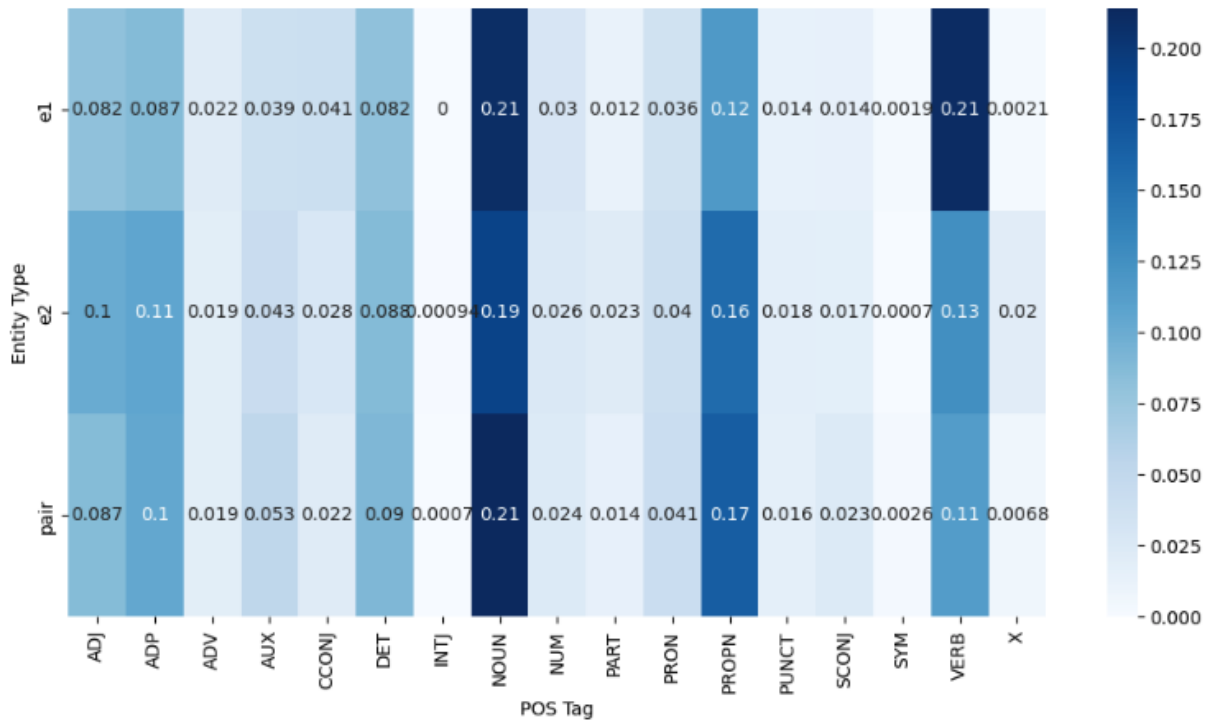


Figure 7: POS feature distribution of TDDAuto

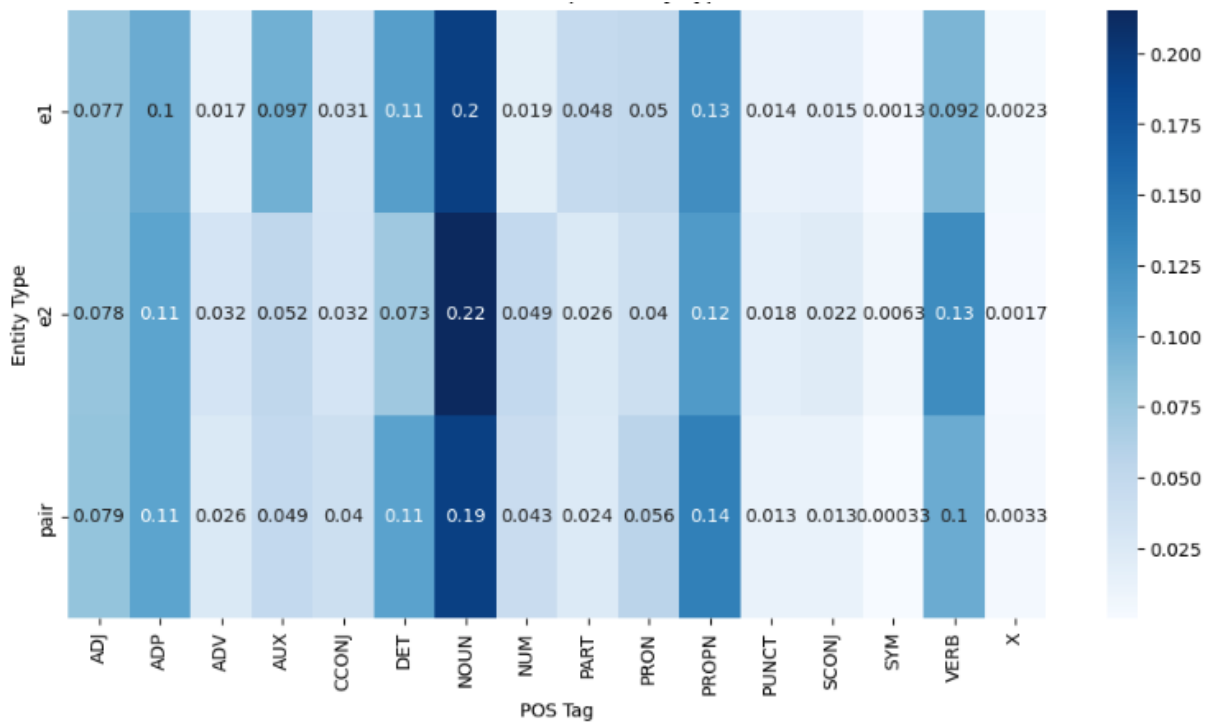


Figure 8: POS feature distribution of TDDman

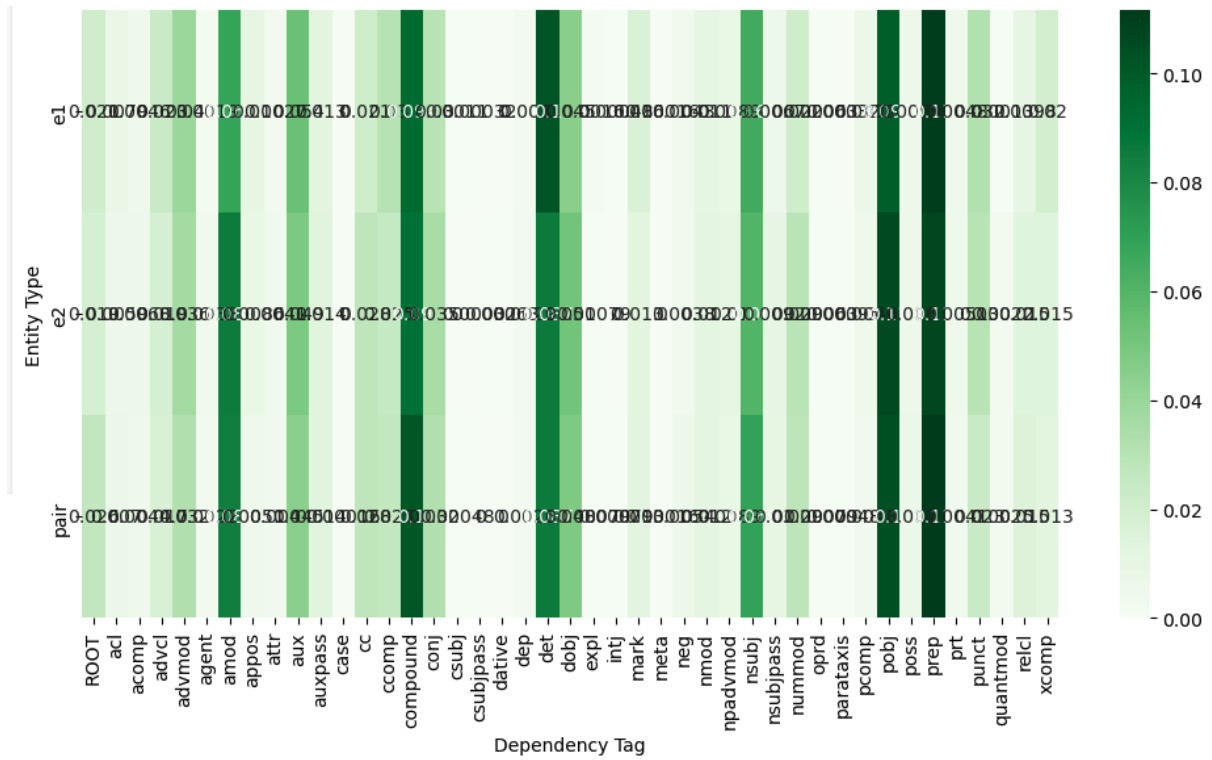


Figure 9: Dependency feature distribution of TimeBank-Dense

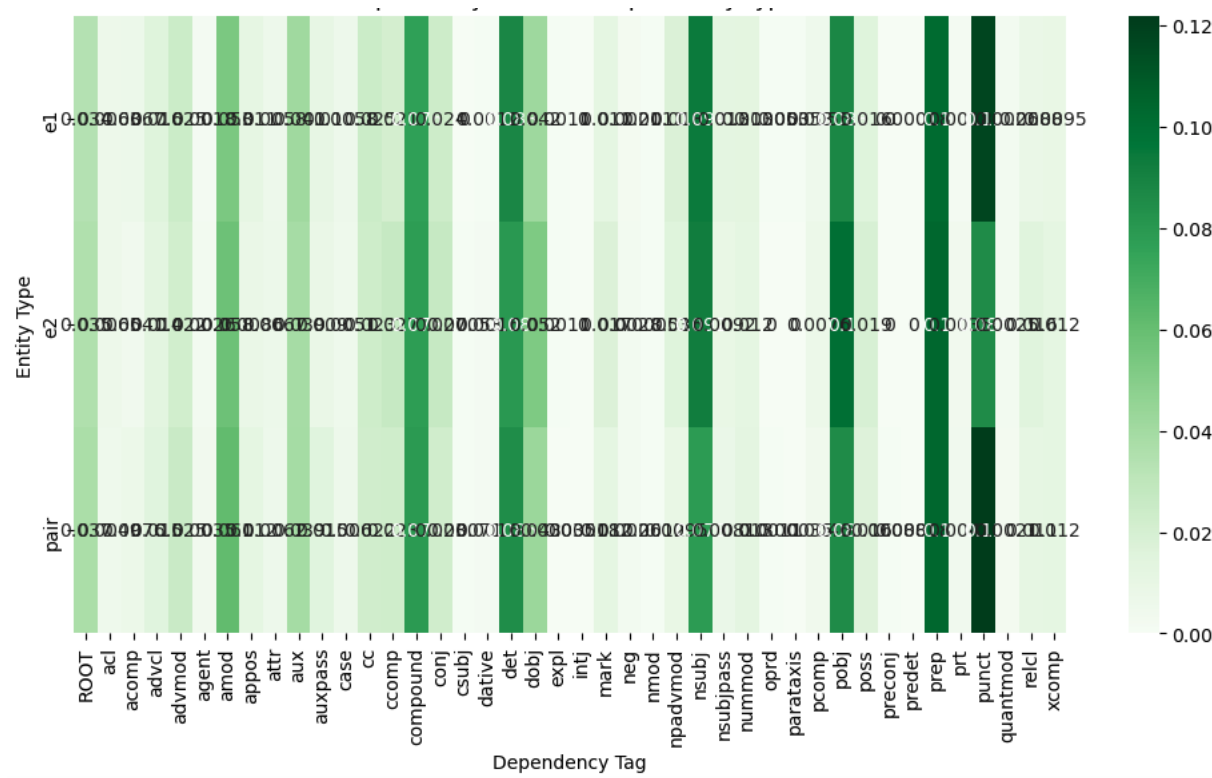


Figure 10: Dependency feature distribution of MATRES

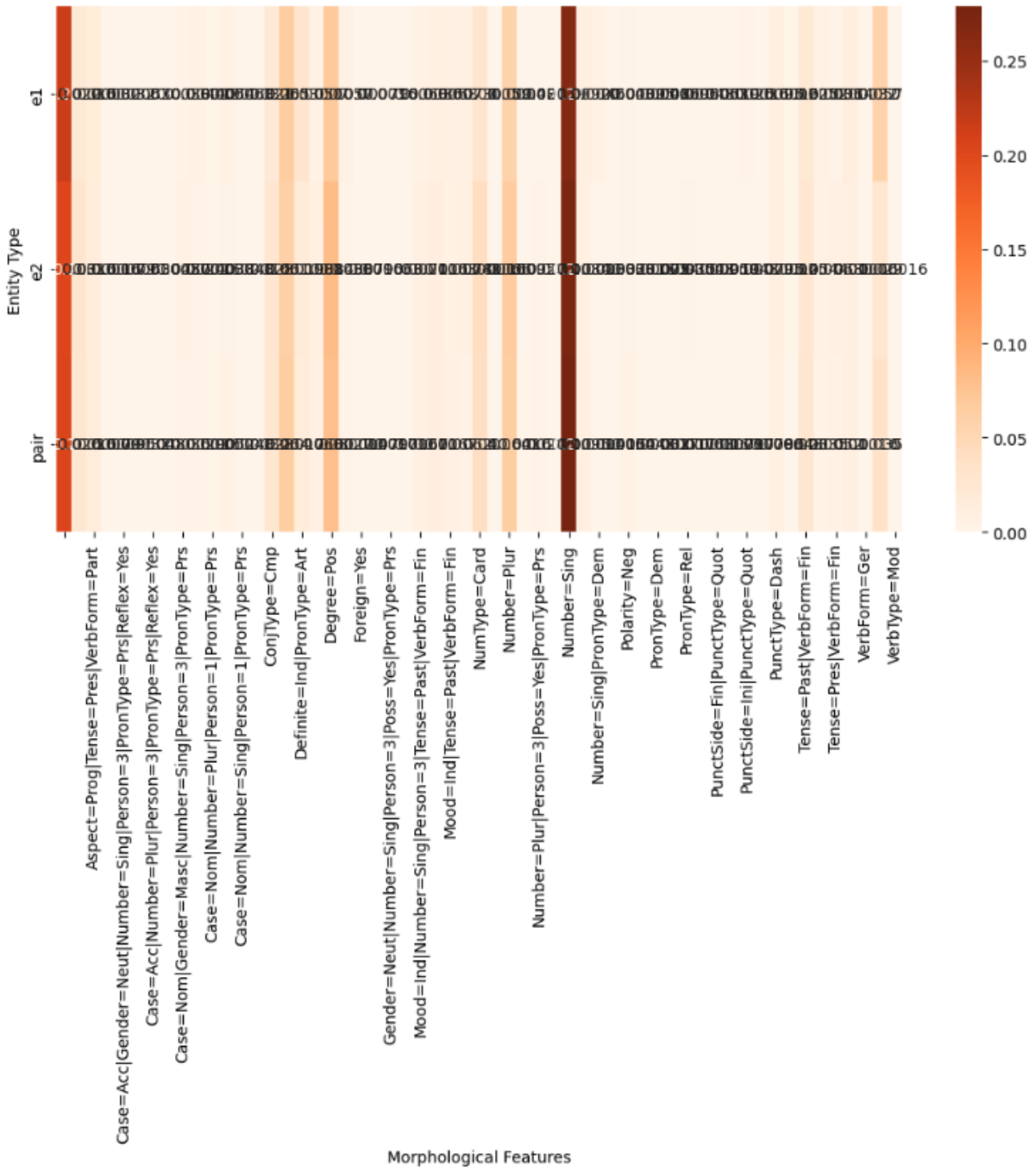


Figure 13: Morphological feature distribution of TimeBank-Dense

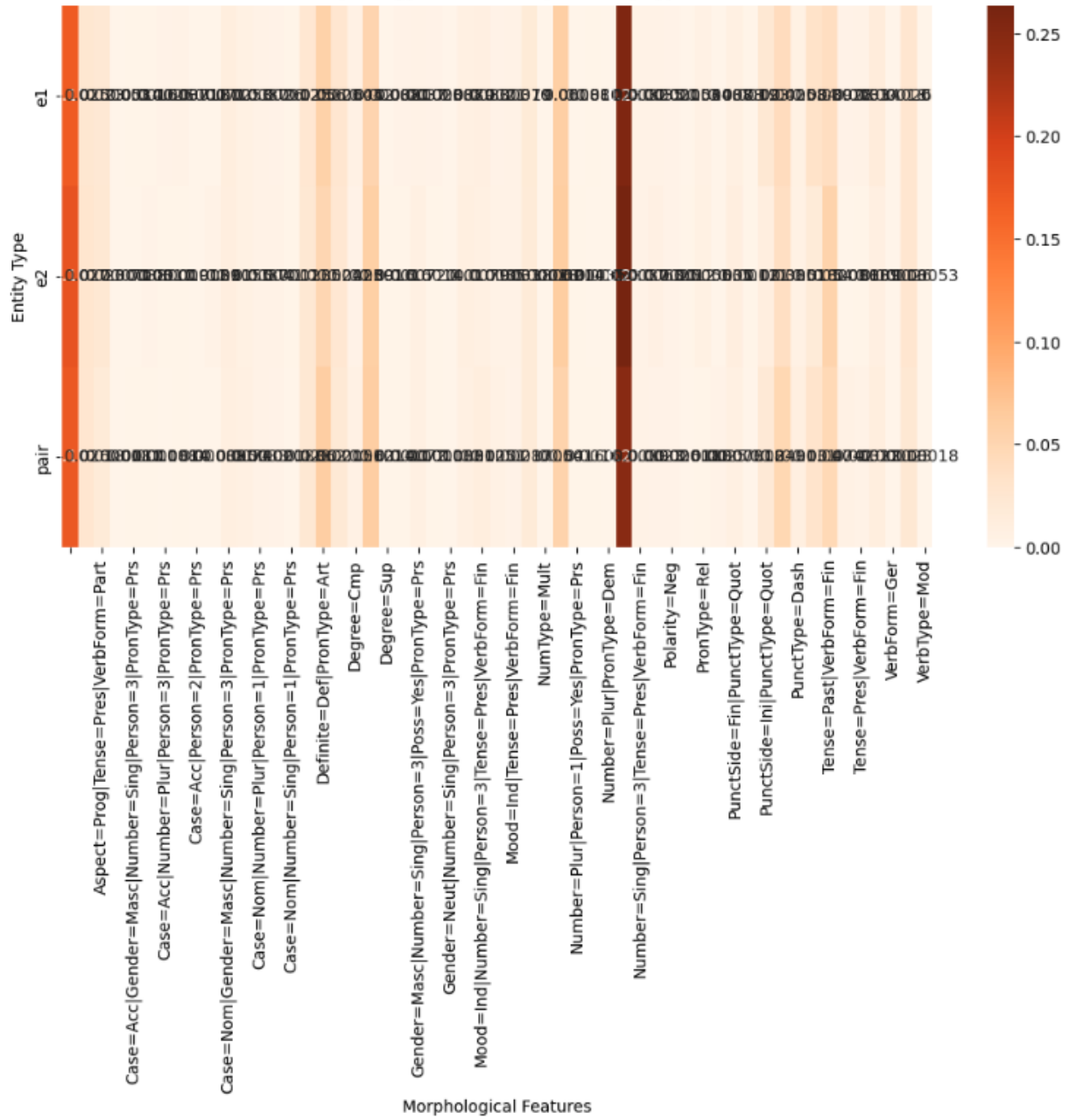


Figure 14: Morphological feature distribution of MATRES

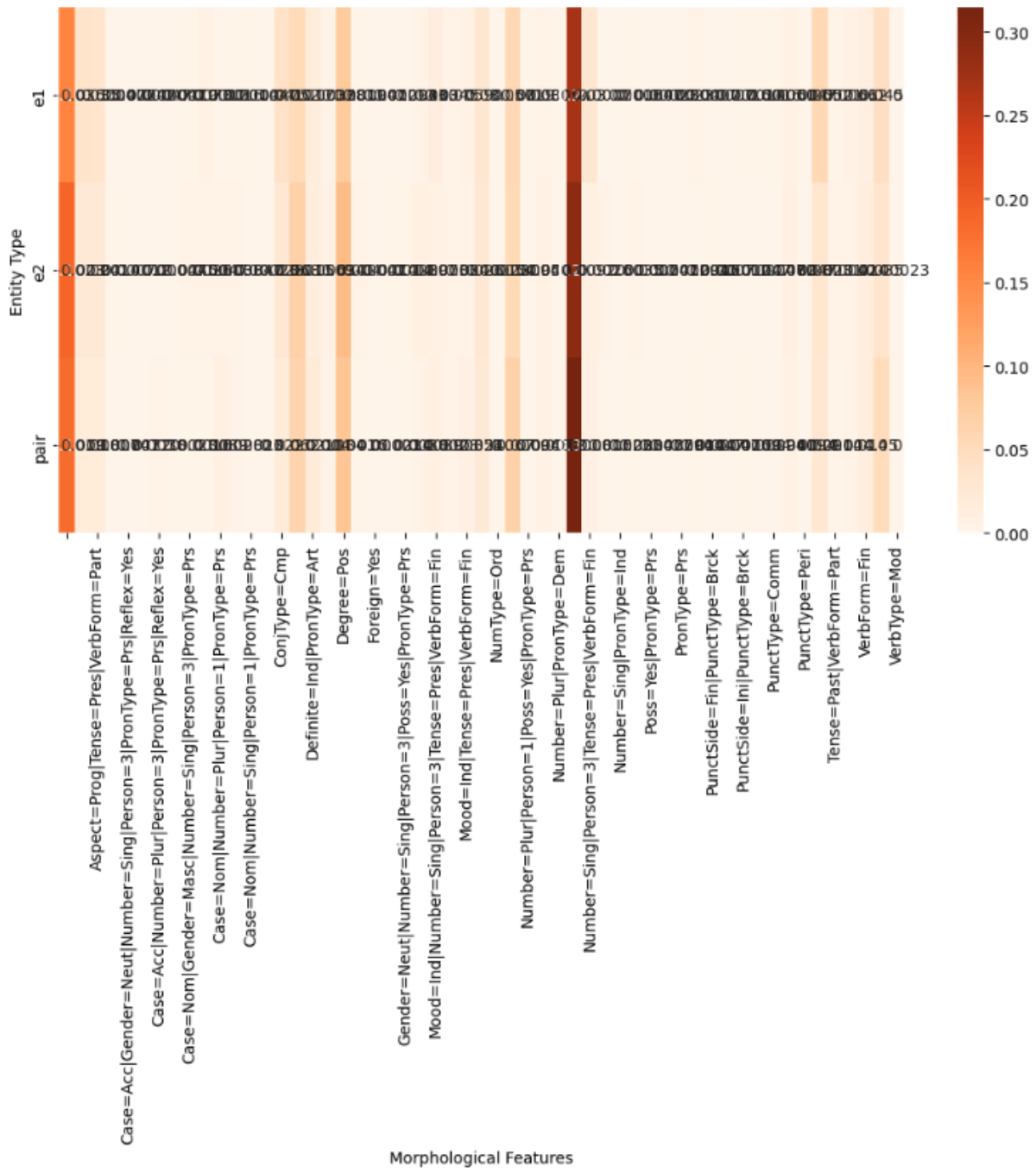


Figure 15: Morphological feature distribution of TDDAuto

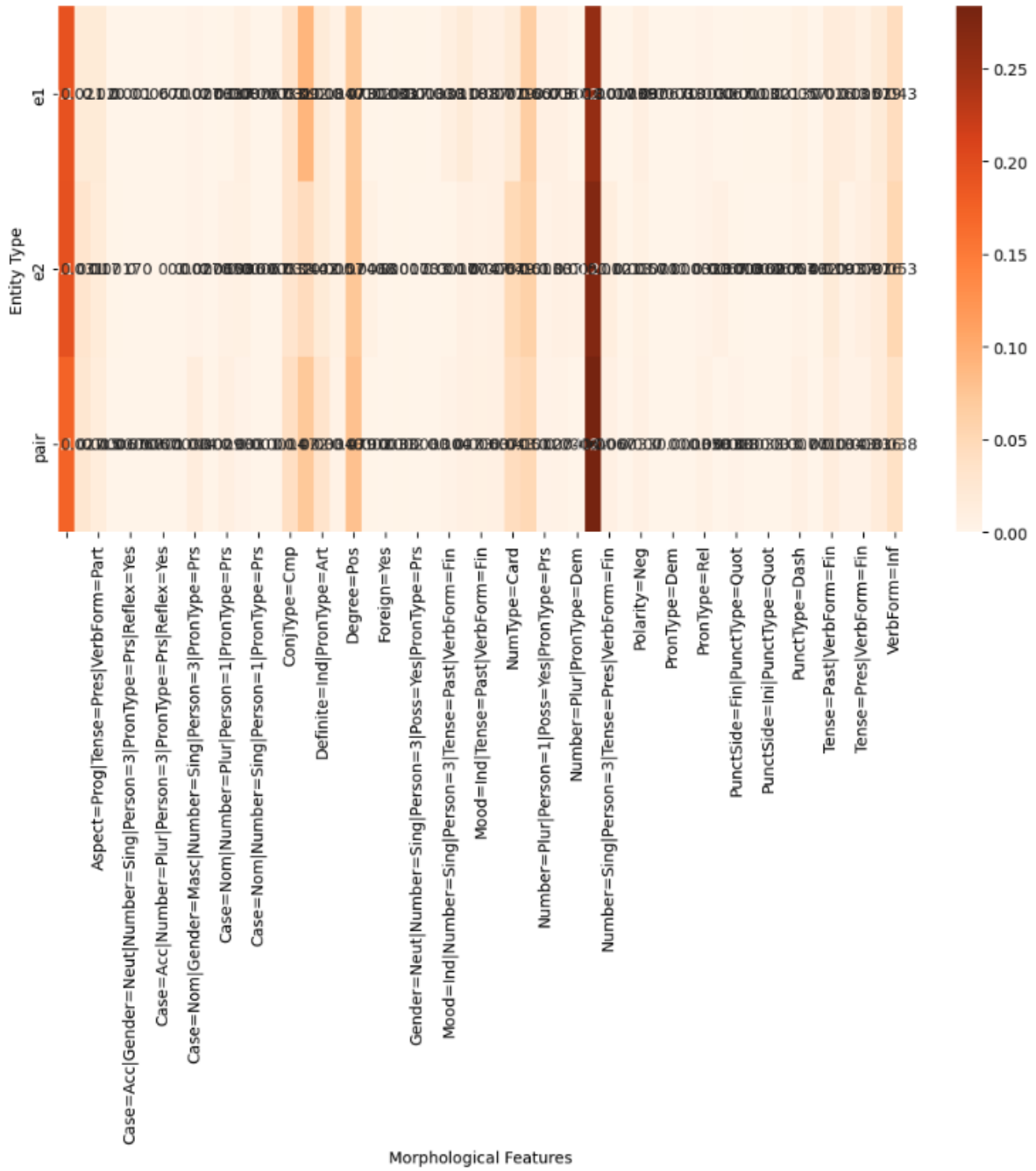


Figure 16: Morphological feature distribution of TDDMan