

Entity or Relation Embeddings?

An Analysis of Encoding Strategies for Relation Extraction

Frank Mtumbuka and Steven Schockaert
Cardiff University, UK
{MtumbukaF, SchockaertS1}@cardiff.ac.uk

Abstract

Existing approaches to relation extraction obtain relation embeddings by concatenating embeddings of the head and tail entities. Despite the popularity of this approach, we find that such representations mostly capture the types of the entities involved, leading to false positives and confusion between relations that involve entities of the same type. Another possibility is to use a prompt with a [MASK] token to directly learn relation embeddings, but this approach tends to perform poorly. We show that this underperformance comes from the fact that information about entity types is insufficiently captured by the [MASK] embeddings. We therefore propose a simple model, which combines such [MASK] embeddings with entity embeddings. Despite its simplicity, our model consistently outperforms the state-of-the-art across several benchmarks, even when the entity embeddings are obtained from a pre-trained entity typing model. We also experiment with a self-supervised pre-training strategy which further improves the results.¹

1 Introduction

Relation extraction consists in identifying the relationships between entities that are expressed in text. It is a fundamental Natural Language Processing (NLP) task, which enables the learning of symbolic representations such as knowledge graphs. While large language models (LLMs) are highly effective in interpreting natural language inputs, the state-of-the-art in relation extraction is still based on fine-tuned language models of the BERT family (Devlin et al., 2019). Moreover, relation extraction is often applied to large document collections, such as corpora of news stories, scientific articles or social media, which means that relying on LLMs is often not feasible in practice, due to their high cost.

¹Our implementation and pre-trained models are available at <https://github.com/fmtumbuka/RelationEmbeddings>

The study of how smaller LMs can be most effectively used for this task thus remains important.

In high-level terms, the standard approach is to fine-tune a BERT-based language model to learn a *relation embedding*, i.e. a vector representation of the relationship that is expressed between two entities in a given sentence, and to train a classifier to predict a discrete *relation label* from this embedding. To obtain these relation embeddings, we cannot simply train a sentence embedding model, since a single sentence may express several relationships. Moreover, we cannot easily identify which tokens in the sentence express the relationship, which makes learning relation embeddings fundamentally different from learning embeddings of text spans. In their seminal work, Baldini Soares et al. (2019) proposed to encapsulate the head and tail entity with special tokens. Consider the following example, where we are interested in extracting the relationship between Paris and France:

The Olympics will take place in [E1] Paris [/E1], the capital of [E2] France [/E2].

The corresponding relation embedding is obtained by concatenating the final-layer embedding of the special tokens [E1] and [E2]. This strategy was found to outperform other alternatives by Baldini Soares et al. (2019), and has remained the most popular approach for learning relation embeddings

The success of this strategy is somewhat surprising: [E1] and [E2] are intuitively designed to represent the head and tail entities, rather than their relationship even though their contextualised representations may still capture the relational context to some extent. In practice, however, knowing the semantic types of the entities often allows us to “guess” the relationship between them, especially if the types are fine-grained. For instance, knowing that the head entity is a capital city and the tail entity is a country, we may reasonably assume that the relationship being expressed is *capital of*.

However, as our analysis in this paper shows, such approaches have at least two important limitations. First, they often struggle to distinguish relations between entities of the same type. Second, they sometimes lead to false positives (e.g. incorrectly assuming that a sentence mentioning a country and a capital city expresses the capital-of relationship). A possible alternative is to add a prompt which includes the [MASK] token, e.g.:

The Olympics will take place in Paris, the capital of France. The relation between Paris and France is [MASK].

We can then fine-tune the language model (LM) such that the contextualised embedding of [MASK] corresponds to a relation embedding. This strategy is popular for zero-shot and few-shot relation extraction (Genest et al., 2022), but is not normally used for the standard supervised relation extraction setting. Our experiments indeed confirm that this approach performs poorly when used alone. Crucially, we find that this strategy struggles because it is not able to accurately characterise the semantic types of the entities, which makes relation prediction much harder. This approach is thus complementary to the entity embedding based strategy: in one case, the model nearly entirely concentrates on entity types, whereas in the other case, it mostly overlooks them. Exploiting this fact, we show that by combining both approaches, we arrive at a simple but highly effective strategy for relation extraction which improves the state-of-the-art in several relation extraction benchmarks. Our main contributions can be summarised as follows:

- We introduce a hybrid strategy which combines the entity embedding and mask based approaches, and we empirically demonstrate its surprising effectiveness.
- We present an analysis of the entity embedding and mask based strategies, showing that the former mostly capture the entity types while the latter do not capture the entity types to a sufficient extent. Inspired by this, we experiment with a variant in which entity embeddings from a pre-trained entity typing model (Mtumbuka and Schockaert, 2023) are used instead of entity embeddings that were trained for relation extraction. Surprisingly, we find that does not deteriorate the results.
- Since the quality of entity and relation embeddings crucially depends on having access to

sufficient training data, we also experiment with a self-supervised pre-training strategy and show that this strategy allows us to further improve the performance of all variants.

2 Related Work

Learning Relation Embeddings The standard approach for relation extraction with LMs uses special tokens to indicate the head and tail entities (also known as the subject and object). Such approaches then predict relation labels from the contextualised representations of these two entities. For instance, the matching-the-blanks model (Baldini Soares et al., 2019) encapsulates the head entity using the special tokens [E1]...[/E1] and the tail entity using separate special tokens [E2]...[/E2]. LUKE (Yamada et al., 2020) simply replaces the entities by the special tokens [HEAD] and [TAIL], omitting the actual entity spans from the input. Wang et al. (2021) encapsulate the head and tail entity with the markers @...@ and #...#, thus avoiding the introduction of new tokens. Some approaches use typed markers, which encode the semantic type of the entity, either as special tokens Zhong and Chen (2021) or by verbalising the entity type as part of the input. It should be noted, however, that entity types are typically not available in practice, which limits the applicability of such approaches.

Another possibility is to append the input with a prompt containing the [MASK] token. This strategy is popular for zero-shot and few-shot relation extraction (Gong and Eldardiry, 2021; Chen et al., 2022; Genest et al., 2022). Rather than training a classifier on top of a relation embedding, the aim is then to compare the contextualised representation of the MASK token with verbalisers, i.e. tokens from the LM’s vocabulary that describe the relationship. This strategy has rarely been considered in the fully supervised setting, with KnowPrompt (Chen et al., 2022) being a notable exception.

Zhong and Chen (2021) already discussed the idea that representing entity types is not sufficient for relation classification. They highlight in particular that jointly training an entity typing and relation extraction system hurts performance, suggesting that both tasks need different kinds of latent representations. We take this idea further, based on the hypothesis that representing relations by concatenating the embeddings of the head and tail entity is inherently limited, even if the entity encoders are specifically trained for relation extraction.

Pre-training Relation Encoders Several approaches have been proposed for pre-training or adapting language models to make them more suitable for the task of relation extraction. The matching-the-blanks model (Baldini Soares et al., 2019) uses entity linking to find sentences that refer to the same entities, and then pre-trains a relation encoder based on the idea that sentences mentioning the same entity pairs are likely to express the same relationship. More recently, variants of this approach based on distant supervision have also been considered. Two sentences are then assumed to have the same relation if their entity pairs are asserted to have the same relation in some knowledge base, a strategy that has a long tradition in relation extraction (Mintz et al., 2009). For instance, Peng et al. (2020) implement this strategy using contrastive learning with the InfoNCE loss. While the aforementioned approaches are focused on fine-tuning a pre-trained LM, the idea of changing the LM model itself has also been explored. For instance, SpanBERT (Joshi et al., 2020) changes the standard masked token prediction task with the aim of learning better span-level representations, while LUKE (Yamada et al., 2020) uses entity linking and distinguished entity tokens to improve the representation of entities.

Relation Extraction with LLMs LLMs such as ChatGPT perform surprisingly poorly on relation extraction benchmarks, and information extraction tasks more generally (Han et al., 2023). Wan et al. (2023) discuss some of the challenges involved in using LLMs for such tasks, which include the difficulty in selecting suitable in-context demonstrations. Peng et al. (2023) make the observation that LLMs with in-context learning (ICL) struggle in particular on specification-heavy tasks, i.e. tasks where even human annotators need to carefully study a non-trivial set of annotation guidelines to correctly solve the task, as is often the case in information extraction. Due to the challenges of using ICL for relation extraction, most approaches involving LLMs use models that can be fine-tuned. For instance, Sainz et al. (2021) proposed a reformulation of relation extraction as a Natural Language Inference (NLI) problem and included experiments with the 1.5B parameter DeBERTa_{XXL} model (He et al., 2021). While this allowed them to improve the state-of-the-art at the time, it should be noted that the NLI based formulation is highly inefficient when a large number of relation labels need to

be considered. It also relies on manually defined verbalisations of the relation labels. Wang et al. (2022a) fine-tuned a 10B parameter model on a range of tasks that can be formulated as triple prediction, including relation extraction. Wadhwa et al. (2023) found that while GPT-3 (Brown et al., 2020) performed poorly when used directly, it was useful for generating chain-of-thought (Wei et al., 2022) explanations. In particular, they showed that by fine-tuning a Flan-T5 (Chung et al., 2022) model on these explanations, the resulting model performed substantially better than when fine-tuning Flan-T5 on the relation extraction task directly. As another strategy for leveraging LLMs indirectly, Xu et al. (2023) use ChatGPT for data generation in few-shot settings. Overall, however, the state-of-the-art in relation extraction, for the fully supervised setting, is still based on fine-tuned models of the BERT family (Devlin et al., 2019). While this might change in future, e.g. with novel prompting techniques or better models, the need for efficient information extraction models means that such smaller models are likely to remain important.

3 Relation Extraction

We consider the standard sentence-level relation extraction setting, where we are given a sentence in which two entities are highlighted, which we refer to as the *head entity* and *tail entity*. The goal is to predict which relationship holds between these two entities, given a pre-defined set of candidate relation labels. We focus on strategies that first learn a relation embedding, which describes the relationship between the two entities in a continuous space, and then use a classifier to predict the actual label based on that relation embedding. In Section 3.1, we first explain the pre-training strategies that we use for learning high-quality relation embeddings. Section 3.2 then describes how the pre-trained relation encoder is fine-tuned for the relation classification task. Finally, in Section 3.3 we explain how relation embeddings can be obtained by concatenating the contextualised embeddings of the head and tail entities or by using a prompt-based strategy with a [MASK] token, among others.

3.1 Pre-training the Relation Encoder

Pre-Training Objective To pre-train relation encoders, we rely on the InfoNCE contrastive loss (van den Oord et al., 2018), which has been found effective for learning relation embeddings (Peng

et al., 2020), and for representation learning in NLP more generally (Gao et al., 2021; Liu et al., 2021; Li et al., 2023; Mtumbuka and Schockaert, 2023). Specifically, let us assume that we have a set S of sentences with designated head and tail entities. For each $s \in S$, we assume that we have access to a set of positive examples P_s , i.e. sentences which express the same relationship as the one expressed in s , and a set of negative examples N_s . Let us write $\phi(s)$ for the relation embedding obtained from sentence s using some encoding strategy. For instance, $\phi(s)$ may be the concatenation of the contextualised representations of the head and tail entity, or it may be the representation of the [MASK] token when a relation prompt is used. Section 3.3 will describe the specific encoding strategies that we consider in our analysis. We train the encoder ϕ using the InfoNCE loss:

$$-\sum_{s \in S} \sum_{p \in P_s} \log \frac{\exp(\cos(\phi(s), \phi(p))/\tau)}{\sum_x \exp(\cos(\phi(s), \phi(x))/\tau)} \quad (1)$$

where the temperature $\tau > 0$ is a hyperparameter, and the summation in the denominator ranges over $x \in N_s \cup \{p\}$. The loss captures the intuition that two sentences expressing the same relationship should have similar relation embeddings. As suggested by previous work (Baldini Soares et al., 2019; Peng et al., 2020) we also include the masked language modelling (MLM) objective during pre-training to prevent catastrophic forgetting. The overall loss is thus given by $\mathcal{L}_{\text{info}} + \mathcal{L}_{\text{MLM}}$, with $\mathcal{L}_{\text{info}}$ the loss in (1) and \mathcal{L}_{MLM} the MLM objective.

Self-Supervised Pre-Training The effectiveness of the pre-training objective crucially depends on the quality and quantity of the available examples. In most cases, we have access to a set of labelled examples, obtained through manual annotation or distant supervision. The positive examples P_s are then simply those examples that have the same label as s . As an alternative, we also experiment with a form of self-supervised pre-training, using coreference chains as a supervision signal. Specifically, we adapt the EnCore strategy from Mtumbuka and Schockaert (2023) for pre-training entity encoders, by proposing a similar strategy for learning relation embeddings. The central idea is that two sentences are likely to express the same relationship if they refer to the same two entities. The matching-the-blank model also relies on this idea, but uses entity linking to identify such sentence pairs. Following Mtumbuka and Schockaert (2023) we select

positive examples from the Gigaword corpus² and we only consider two entities to be co-referring if they are identified as such by two separate off-the-shelf coreference systems: the *Explosion AI* system Coreferee v1.3.1³ and the *AllenNLP* coreference model⁴. This use of two coreference systems was found to reduce the number of false positives because of spurious coreference links, given that state-of-the-art coreference resolution systems are still far from perfect. Clearly, the fact that two sentences mention the same entities does not guarantee that the sentences actually express the same relationship, which is a common limitation of self-supervised strategies. Note, however, that this issue is somewhat mitigated because we only consider sentences from the same news story.

3.2 Relation Classification

Given a sentence s with designated head and tail entities, we use the pre-trained relation encoder to obtain an embedding $\phi(s)$ that captures the relationship between these entities. Now consider the problem of classifying this relationship, using the labels from some set $\{l_1, \dots, l_m\}$. Following standard practice (Zhong and Chen, 2021; Zhou and Chen, 2022), we use a feedforward network with one hidden layer and ReLU activation, i.e. predictions are made as follows:

$$\mathbf{h} = \text{ReLU}(\mathbf{A}_1 \phi(s) + \mathbf{b}_1) \\ (p_1, \dots, p_m) = \text{softmax}(\mathbf{A}_2 \mathbf{h} + \mathbf{b}_2)$$

where p_i is interpreted as the probability that l_i is the correct label, \mathbf{A}_1 and \mathbf{A}_2 are matrices, and \mathbf{b}_1 and \mathbf{b}_2 are bias terms. The label classifier is trained using cross-entropy. We also fine-tune the pre-trained relation encoder during this step. The dimension of the hidden representations \mathbf{h} is set to be the same as that of the corresponding encoder. For instance, for all experiments with BERT-base, we set the hidden layer to 768 dimensions.

3.3 Encoding Strategies

We now discuss the considered strategies for obtaining relation embeddings. Suppose we are interested in the relationship between the entities $\langle h \rangle$ and $\langle t \rangle$ in sentence s . We first create an annotated version of sentence s , where (i) $\langle h \rangle$ is encapsulated

²<https://catalog.ldc.upenn.edu/LDC2003T05>

³<https://github.com/explosion/coreferee>

⁴<https://demo.allennlp.org/coreference-resolution>

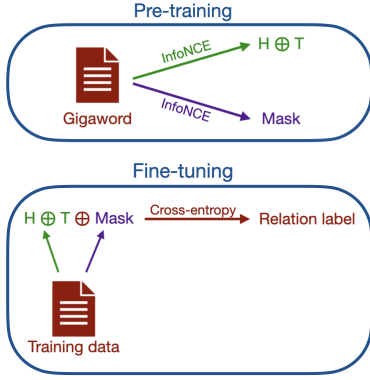


Figure 1: Illustration of the [H,T]+Mask strategy.

with the special tokens [E1]...[/E1] and $\langle t \rangle$ is encapsulated with the special tokens [E2]...[/E2], and we append the phrase “The relation between $\langle h \rangle$ and $\langle t \rangle$ is [MASK]”. For instance:

The Olympics will take place in [E1] Paris [/E1], the capital of [E2] France [/E2]. The relation between Paris and France is [MASK].

For some of our strategies the [MASK] token will not be used, while another strategy only uses the [MASK] token. However, we use the same annotated sentence in all cases, as this allows for the most direct comparison. First, we consider the following basic approaches:

[H,T] We define $\phi(s)$ as the concatenation of the final-layer embeddings of the tokens [E1] and [E2], following [Baldini Soares et al. \(2019\)](#).

Mask We define $\phi(s)$ as the final-layer representation of the [MASK] token.

[H,T,Mask] We define $\phi(s)$ as the concatenation of the final-layer embeddings of the tokens [E1], [E2] and [MASK].

In each case, we pre-train the relation encoder as explained in Section 3.1 and then train a classifier as explained in Section 3.2.

Hybrid Strategy We also consider a hybrid approach, where we pre-train the relation encoder using a loss of the form $\mathcal{L}_{\text{info}}^1 + \mathcal{L}_{\text{info}}^2 + \mathcal{L}_{\text{MLM}}$. $\mathcal{L}_{\text{info}}^1$ and $\mathcal{L}_{\text{info}}^2$ both relate to the InfoNCE loss (1) but refer to different representations: $\mathcal{L}_{\text{info}}^1$ uses the [H,T] representation while $\mathcal{L}_{\text{info}}^2$ uses the Mask representation. Note that we fine-tune a single language model, i.e. the [H,T] and Mask representations are obtained with the same encoder. After the pre-training step, when training the classifier,

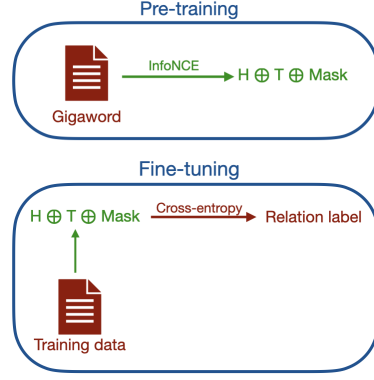


Figure 2: Illustration of the [H,T,Mask] strategy.

we concatenate the two representations (i.e. the embeddings of the [E1], [E2] and [MASK] tokens). We will refer to this strategy as **[H,T]+Mask**. Note that the only difference with [H,T,Mask] lies in how the encoder is pre-trained. The underlying motivation comes from the idea that the [MASK] token may be largely ignored when using the [H,T,Mask] strategy, since it may be easier to learn meaningful representations of the entities than to learn meaningful relation embeddings. The hybrid training strategy [H,T]+Mask avoids this issue, by forcing the [MASK] token to capture a meaningful relation embedding, before combining this representation with the contextualised entity embeddings. Figures 1 and 2 illustrate the difference between the [H,T,Mask] and [H,T]+Mask strategies.

Pre-trained Entity Embeddings Our main hypothesis is that the [H,T] strategy focuses on learning the semantic types of the head and tail entities. To test this hypothesis, we consider a variant in which we use an entity embedding model instead of pre-training a relation encoder using the [H,T] strategy. In particular, we rely on EnCore ([Mtumbuka and Schockaert, 2023](#)) as the pre-trained entity embedding model. The EnCore embeddings essentially capture the semantic types of the entities (but without reference to any particular set of type labels). Crucially, no relational knowledge is used during the EnCore training process. We consider the following variants:

EnCore No pre-training is used. We directly train a classifier on the concatenation of the EnCore embeddings of the head and tail entities.

EnCore+Mask We pre-train an encoder using the Mask strategy. Then we train the classifier on the concatenation of the [MASK] token and the EnCore entity embeddings.

EnCore+[H,T]+Mask We use the same hybrid pre-training as the [H,T]+Mask strategy. The classifier is then trained on the concatenation of the [H,T]+Mask representation and the EnCore entity embeddings.

Note that the EnCore model itself is not fine-tuned. This ensures that the embeddings provided by this model remain focused on entity types.

4 Experiments

We compare the effectiveness of the considered relation embedding strategies on a number of standard relation extraction benchmarks. Our main hypothesis is that the common [H,T] strategy essentially leads models to focus on the semantic types of the head and tail entity. We are thus interested in comparing the [H,T] and EnCore strategies. Furthermore, we hypothesise that the information captured by the Mask embeddings is complementary to that captured by the [H,T] embeddings. Accordingly, we are interested in comparing [H,T] with [H,T,Mask] and [H,T]+Mask.

4.1 Experimental Setup

Benchmarks We evaluate on five standard benchmarks. First, we use TACRED (Zhang et al., 2017), a popular relation extraction benchmark. Two revisions of this dataset have been proposed, both of which are aimed at addressing problems with noisy annotations. In particular, TACREV (Alt et al., 2020) was obtained by re-annotating the 5000 most challenging instances from the development and test sets. Specifically, to select these instances, the authors looked at how often models disagreed with each other and with the ground truth. ReTACRED (Stoica et al., 2021) was obtained by re-annotating the entire dataset. Following tradition, we report results on all three variants in terms of F1 score. Next, we evaluate on a distantly supervised dataset that was introduced by Sorokin and Gurevych (2017) by aligning Wikipedia and Wikidata (Wiki-WD). Finally, we also consider the distantly supervised dataset that was introduced by Riedel et al. (2010) by aligning articles from the New York Times with Freebase (NYT-FB). The NYT-FB dataset does not have an explicit validation set. As a result, we keep 10% of the training set as a validation set and train on the remaining 90%. Following the tradition from previous work, we report the results on Wiki-WD in terms of F1

Dataset	# Class	Train	Dev.	Test
TACRED	42	68.1K	22.6K	15.5K
TACREV	42	68.1K	22.6K	15.5K
ReTACRED	40	58.4K	19.5K	13.4K
Wiki-WD	353	372.1K	123.8K	360.3K
NYT-FB	53	455.8K	-	172.4K

Table 1: Overview of the considered benchmarks, showing the number of distinct relation classes, and the number of annotated mentions in the training, development and test sets.

score and the results on NYT-FB in terms of precision at 10 (P@10) and 30 (P@30), averaged across all relation labels. Table 1 summarises the main characteristics of the considered datasets.

Baselines FOR TACRED and its variants, we consider a number of recent baselines. First, we include a comparison with SpanBERT (Joshi et al., 2020) and LUKE (Yamada et al., 2020). We furthermore compare with KnowBERT (Chen et al., 2022), which also uses a prompt with the [MASK] token. They improve on the standard Mask strategy, among others, by incorporating predicted entity types, and are thus a natural baseline for our methods. Finally, we consider the Typed Marker strategy from Zhou and Chen (2022), and the Curriculum Learning variant from Park and Kim (2021). Note, however, that these last two methods are not directly comparable with our methods, as they rely on the gold entity type labels that are provided as part of the TACRED dataset. We do not use these labels for our models since such information is typically not available in practice. For Wiki-WD and NYT-FB, we compare with RECON (Bastos et al., 2021) and KGPool (Nadgeri et al., 2021). Note, however, that these state-of-the-art methods are again not directly comparable with our methods. They are focused on modelling and exploiting knowledge from the Wikidata and Freebase knowledge graphs. We do not take such information into account, as our focus is on comparing different encoding strategies for learning relation embeddings.

Training Strategies Unless stated otherwise, the relation encoder is first pre-trained using the strategy from Section 3.1, before being fine-tuned, as explained in Section 3.2. In our default setting, we use the same training set for pre-training and fine-tuning the relation encoder. For TACRED, TACREV and ReTACRED, where the training set is comparatively smaller, we also experiment with a

	TACRED	TACREV	Re-TACRED
<i>Baselines</i>			
SpanBERT	70.8	78.0*	85.3 [†]
LUKE	72.7	80.6 [◦]	90.3 [◦]
KnowPrompt	72.4	81.4	90.9
Typed Marker	74.6	83.2	91.1
Curriculum Learning	75.0	-	91.4
<i>Standard pre-training</i>			
Mask	23.3	22.9	23.2
[H,T]	61.4	63.6	72.3
[H,T,Mask]	73.0	73.8	81.9
[H,T]+Mask	78.5	84.6	91.9
EnCore+Mask	78.9	83.9	91.8
EnCore+[H,T]+Mask	78.1	84.1	91.8
<i>Gigaword pre-training</i>			
Mask	24.6	23.5	38.2
[H,T]	63.2	66.3	79.7
[H,T,Mask]	74.4	75.1	82.9
[H,T]+Mask	78.5	83.6	92.7
EnCore+Mask	79.1	84.9	93.5
EnCore+[H,T]+Mask	79.0	84.4	93.2
<i>No pre-training</i>			
[H,T,Mask]	55.2	54.3	60.4
EnCore	69.8	76.1	80.4

Table 2: Comparison of different relation embedding strategies, in terms of F1 (%). Results marked with * were taken from Alt et al. (2020), those marked with † were taken from Stoica et al. (2021), and those marked with ◦ were taken from Zhou and Chen (2022). All other baseline results were taken from the original papers. Our models are initialised from RoBERTa-large, which is the same for all baselines except for SpanBERT, which uses BERT-large.

variant where we instead use the Gigaword corpus for pre-training the relation encoder, following the self-supervision strategy from Section 3.1. Finally, we also report results where the pre-training step is omitted and we directly train the relation encoder using the fine-tuning strategy from Section 3.2. For our main experiments, we use roberta-large⁵ to initialise the relation encoder.

4.2 Results

Table 2 summarises the results for TACRED, TACREV and Re-TACRED. Let us first consider the standard pre-training results (i.e. where the models are pre-trained on the standard training set).

⁵https://huggingface.co/docs/transformers/model_doc/roberta

	Wiki-WD			NYT-FB	
	P	R	F1	P@10	P@30
<i>Baselines</i>					
RECON	87.2	87.2	87.2	87.5	74.1
KGPool	88.6	88.6	88.6	92.3	86.7
<i>Standard pre-training</i>					
Mask	50.3	48.7	48.9	52.3	49.7
[H,T]	75.9	74.7	75.2	79.6	78.5
[H,T,Mask]	82.1	80.8	81.7	88.4	86.1
[H,T]+Mask	89.8	89.3	89.6	94.7	93.1
EnCore+Mask	89.4	88.8	89.1	94.9	93.0
EnCore+[H,T]+Mask	89.8	88.6	88.9	94.7	92.9
<i>No pre-training</i>					
[H,T,Mask]	56.1	54.7	55.1	58.1	56.9
EnCore	80.8	79.2	79.8	86.1	84.7

Table 3: Comparison of different relation embedding strategies. Baseline results were obtained from the original papers. Our models are initialised from RoBERTa-large.

A number of observations stand out. First, the Mask strategy on its own performs poorly. Second, there is clear evidence that the [H,T] and Mask strategies are complementary: both [H,T,Mask] and [H,T]+Mask substantially outperform [H,T] and Mask on their own. We can furthermore see that [H,T]+Mask outperforms [H,T,Mask]. The combined pre-training strategy which is used by [H,T,Mask] thus indeed seems to largely fail to learn meaningful [MASK] embeddings. Finally, the EnCore+Mask strategy matches or even outperforms [H,T]+Mask. This is surprising, given that the EnCore embeddings are specifically trained to capture semantic types and are not fine-tuned on the relation extraction datasets. The strong performance of EnCore+Mask thus clearly supports the idea that the [H,T] representations mostly capture the entity types, rather than the relationship itself. We can similarly see that EnCore+[H,T]+Mask does not generally improve on EnCore+Mask, which further supports this idea.

Pre-training on Gigaword leads to clear and consistent improvements, compared to standard pre-training. While the fact that pre-training on external datasets can bring benefits is in itself unsurprising, this self-supervision strategy based on coreference chains was not previously tested for relation extraction. It offers a convenient way to improve relation extraction systems, since it does not rely on an entity linked corpus. Comparing the performance of the different configurations, after pre-training

Sentence	Label
Asia Bibi was sentenced to hang in Pakistan’s central province of Punjab earlier this month after being accused of insulting the Prophet Mohammed in 2009.	Gold: no relation Mask: per:origin [H,T]: per:state or provinces of residence [H,T]+Mask: no relation
In October, she filed a complaint with the police in Rio saying he had kidnapped her and tried to threaten her into having an abortion.	Gold: no relation Mask: per:origin [H,T]: per:cities of residence [H,T]+Mask: no relation
Benjamin Chertoff is the Editor in Chief of Popular Mechanics magazine as well as the cousin of the Director of Homeland Security , Michael Chertoff.	Gold: no relation Mask: per:employee of [H,T]: per:employee of [H,T]+Mask: no relation
WASHINGTON – The National Restaurant Association gave \$35,000 – a year’s salary – in severance pay to a female staff member in the late 1990s after an encounter with Herman Cain , its chief executive at the time, made her uncomfortable working there, three people with direct knowledge of the payment said on Tuesday.	Gold: org:top members/employees Mask: per:employee of [H,T]: no relation [H,T]+Mask: org:top members/employees
“From January 1, I, Charles Ble Goude and the youth of Ivory Coast are going to liberate the Golf Hotel with our bare hands,” the political showman turned minister declared Wednesday, to a cheering crowd of hardline supporters.	Gold: per:title Mask: per:employee of [H,T]: no relation [H,T]+Mask: per:title

Table 4: Comparison of the TACREV test set predictions from the Mask, [H, T] and [H, T] + Mask models that were initialised using RoBERTa-large.

on Gigaword, we see the same patterns as with standard pre-training. The Encore+Mask strategy emerges as the best model overall, which in particular confirms the usefulness of combining entity embeddings information with relation embeddings.

We can also see that forgoing pre-training altogether has a detrimental effect. A direct comparison with the baselines is difficult, as the two strongest baselines use additional information (i.e. the gold entity type labels). Nonetheless, our best configurations consistently outperform the state-of-the-art methods, with the improvements being most pronounced for TACRED.

Table 3 summarises the results we obtained for Wiki-WD and NYT-FB. The main patterns are consistent with the results from Table 2. For instance, we can again see that Mask on its own performs poorly and that [H,T]+Mask outperforms both Mask and [H,T] by a considerable margin. We furthermore again see that the pre-trained entity embeddings from EnCore can serve the same purpose as the fine-tuned [H,T] embeddings. The best configurations outperform the baselines, with the improvements on NYT-FB being particularly clear. However, these methods are not directly comparable as they focus on different information.

Further analysis of our results can be found in the appendix. Among others, we show that the conclusions from this section remain valid when other

language models than RoBERTa-large are used as the encoder. We also present a detailed error analysis to support our claims about the limitations of the [H,T] and Mask strategies, an analysis of the models in a setting with limited training data, and an analysis of the impact of the dimensionality of the entity and relation embeddings.

4.3 Qualitative Analysis

In Table 4, we compare predictions of the Mask, [H,T] and [H,T]+Mask models for the TACREV test set. The first three cases illustrate how the [H,T] model often overly relies on the semantic types of the entities, without fully taking into account the actual sentence context, a problem known as entity bias (Wang et al., 2022b). In particular, as the first two examples illustrate, given a person and a place, the [H,T] model frequently predicts that the individual is a resident of that place, even if there is no relationship expressed in the given context. In contrast, [H,T]+Mask correctly predicts *no relation* in these cases. The third example shows a similar issue, which arises when the sentence refers to a person and an organisation. In this case, the [H,T] model incorrectly predicts that the person is employed by that organisation. As the fourth example illustrates, however, the opposite situation also arises, where the [H,T] model incorrectly predicts *no relation*. Furthermore, several examples illus-

trate how the Mask model struggles because it does not adequately capture the semantic types of the entities. For instance, in the first two examples, the model predicts *per:origin* despite the fact that the tail entity is not a country. The issue is most clearly illustrated by the fifth example, where the Mask model predicts *employee of*, despite the tail entity not being an organisation. Overall, these examples support the view that the [H,T] model focuses too much on modelling the entity types, which is not always sufficient, while conversely, Mask struggles because it does not sufficiently take entity types into account. We include further examples in the appendix, which further support our findings.

5 Conclusions

The primary aim of this paper was to analyse two different strategies for training relation encoders. On the one hand, most work in supervised relation extraction relies on contextualised embeddings of the head and tail entity for predicting relationships. On the other hand, prompt-based strategies can be used to obtain embeddings that represent the relationship itself. The latter strategy is arguably more intuitive, but we found it to perform poorly in practice. Rather than suggesting that such relation embeddings are not useful, however, we found that they capture information that is highly complementary to what is captured by contextualised entity embeddings. Indeed, we considered a hybrid strategy, which substantially outperforms either of the two individual strategies, allowing us to improve the state of the art in each of the five considered benchmarks. Remarkably, we found that this remains true if we use entity embeddings from an off-the-shelf entity encoder. Finally, as a secondary contribution, we also found that coreference chains, which were used for training entity encoders in a self-supervised way by [Mtumbuka and Schockaert \(2023\)](#), can be successfully leveraged for self-supervised training of relation encoders.

Limitations

Our analysis in this paper was limited to the English language, and we only considered the setting of fully supervised sentence-level relation extraction. Since our focus was on learning representations (i.e. relation embeddings), it is not straightforward to transfer our findings to the zero-shot relation extraction setting (as we do not attempt to model the relation labels). We did not consider the

use of LLMs for relation extraction in this paper. On the one hand, this is due to the fact that applying LLMs to this setting is not straightforward, especially when the set of candidate relation labels is large. Progress in this area has indeed been slow, as we highlighted in the related work section. Moreover, while LLMs can be used to extract symbolic representations (e.g. knowledge graph triples), they are often less suitable for learning embeddings. Traditionally, relation embeddings have primarily been used as an intermediate representation, before relation labels were predicted, and from this perspective, we may wonder whether relation embeddings are still needed in the LLM era. However, beyond acting as an intermediate representation, embeddings have a number of important advantages. They can, in principle, capture much more subtle distinctions than is possible with pre-defined discrete relation labels. As such, they are more suitable for modelling relational similarity (i.e. analogy), for instance.

Acknowledgements This research was supported by EPSRC grant EP/W003309/1 and undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of the Cardiff Supercomputing Facility and the HPC Wales and Supercomputing Wales (SCW) projects. We acknowledge the support of the latter, which is part-funded by the European Regional Development Fund (ERDF) via the Welsh Government.

References

- Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. [TACRED revisited: A thorough evaluation of the TACRED relation extraction task](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, 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.
- Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. [RECON: Relation extraction using knowledge graph context in a graph neural network](#). In *Proceedings of the Web Confer-*

- ence 2021, WWW '21, page 1673–1685, New York, NY, USA. Association for Computing Machinery.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. [KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction](#). In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *CoRR*, abs/2210.11416.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Laurent-Walter Goix. 2022. [PromptORE - A novel approach towards fully unsupervised relation extraction](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 561–571. ACM.
- Jiaying Gong and Hoda Eldardiry. 2021. [Prompt-based zero-shot relation classification with semantic knowledge augmentation](#). *CoRR*, abs/2112.04539.
- Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. [Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors](#). *CoRR*, abs/2305.14450.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: decoding-enhanced bert with disentangled attention](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [SpanBERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Na Li, Hanane Kteich, Zied Bouraoui, and Steven Schockaert. 2023. [Distilling semantic concept embeddings from contrastively fine-tuned language models](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 216–226. ACM.
- Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021. [MirrorWiC: On eliciting word-in-context representations from pretrained language models](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. [Summarization as indirect supervision for relation extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. [Distant supervision for relation extraction without labeled data](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.
- Frank Mtumbuka and Steven Schockaert. 2023. [EnCore: Pre-training entity encoders using coreference chains](#). *CoRR*, abs/2305.12924.
- Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang, Johannes Hoffart, Saeedeh

- Shekarpour, and Vijay Saraswat. 2021. [KGPool: Dynamic knowledge graph context selection for relation extraction](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, Online. Association for Computational Linguistics.
- Seongsik Park and Harksoo Kim. 2021. [Improving sentence-level relation extraction through curriculum learning](#). *CoRR*, abs/2107.09332.
- Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. [Learning from Context or Names? An Empirical Study on Neural Relation Extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.
- Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. [When does in-context learning fall short and why? A study on specification-heavy tasks](#). *CoRR*, abs/2311.08993.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Oscar Sainz, Oier Lopez de Lacalle, Gorika Labaka, Ander Barrena, and Eneko Agirre. 2021. [Label verbalization and entailment for effective zero and few-shot relation extraction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Daniil Sorokin and Iryna Gurevych. 2017. [Context-aware representations for knowledge base relation extraction](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark. Association for Computational Linguistics.
- George Stoica, Emmanouil Antonios Platanios, and Barnabas Poczos. 2021. [Re-TACRED: Addressing shortcomings of the TACRED dataset](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13843–13850.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. [Revisiting relation extraction in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.
- Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. [GPT-RE: In-context learning for relation extraction using large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.
- Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022a. [DeepStruct: Pre-training of language models for structure prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.
- Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. [K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.
- Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022b. [Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. [How to unleash the power of large language models for few-shot relation extraction?](#) In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing, SustaiNLP 2023, Toronto, Canada (Hybrid), July 13, 2023*, pages 190–200. Association for Computational Linguistics.
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. [An improved baseline for sentence-level relation extraction](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

A Training Details

We pre-train the model for 25 epochs and select the checkpoint with the minimum validation loss. For the fine-tuning step, we similarly train the model for 25 epochs and select the best checkpoint based on the validation set. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $5e - 4$ and a weight decay λ of $1e - 5$. The temperature τ in the contrastive loss was set to 0.05.

B Analysis

In this section we provide some further analysis of the different relation embeddings. In particular, we make the following observations:

- By analysing model outputs, we find that the “no relation” label (from TACRED) is the most difficult label for all variants, where models are particularly prone to predicting some relationship even if none is expressed in the sentence. The [H,T]+Mask variant suffers considerably less from this problem, which helps to explain its outperformance over [H,T].
- In a qualitative analysis of the model outputs, we provide further evidence that [H,T] focuses too much on the entity types. In particular, the model frequently predicts a relationship that holds for entities of the same type but is not expressed in the sentence.
- We show that the predictions of the masked language model for the [MASK] token are largely meaningless, even though the contextualised representation of this token is useful for relation classification. This is expected, given that the model is not trained to produce meaningful verbalisations of the relationship.

Confusion Matrices In Figures 3, 4 and 5, we show the confusion matrix for 5 randomly sampled relations and the “no relation” label, for the TACREV test set. For this analysis, we focus on the Mask, [H,T] and [H,T]+Mask models. We refer to any label predictions that fall outside of the six sampled classes as the “other” class. We can clearly observe that the “no relation” label is the most difficult label for all three models: most errors occur when the models predict a relationship while none exists, or vice versa. We can also see that the [H,T]+Mask model regularly outperforms [H,T], which in turn outperforms the Mask model when it comes to modelling the “no relation” label.

Error Comparison In Figure 6, we present the analysis of model performance that highlights instances where one model outperforms the other. Specifically, we focus on the [H,T] and [H,T]+mask models, showing how often one model makes a mistake while the other gets it right, for the TACREV test set. Compatible with our findings from the confusion matrices, the no “relation” label proves challenging for both methods. There are several instances where [H,T] makes a mistake while [H,T]+Mask does not, and vice versa, although [H,T]+Mask overall performs better. When we look at the other relation labels, however, we can see a clear pattern, as there are only very few instances where [H,T]+Mask makes a mistake without [H,T] also making the same mistake. This shows that the improvement of [H,T]+Mask over [H,T] is highly consistent.

Qualitative Analysis of Model Outputs Overall, we find that around 70% of the errors of the [H,T] model are “no relation” misclassifications, where most of the remaining mistakes arise because the model confuses relations between entities of the same type, including:

- family relationships such as “per:siblings” and “per:children”;
- relationships linking people to geographic regions, such as “per:state or province of death” and “per:state or province of residence”;
- relationships linking people to organisations, such as “org:founded by” and “org:top members/employees”.

We provide several examples of both types of errors. In particular, Table 5 shows cases where [H,T]

mistakenly predicts the “no relation” label, while Table 6 focuses on examples where the [H,T] model confuses the target relation with a relation between entities of the same type.

The Mask model performs worse overall, and there is less of a pattern in the types of errors it makes. Similar to the other models, it is also prone to incorrectly predicting “no relation”. Several examples of this can be seen in Tables 5 and 6. One particular weakness of Mask is that it sometimes fails to correctly predict the direction of a relationship, confusing the target relation with its inverse. This can be seen, for instance, in the last example of Table 6, where Mask confuses “per:children” with “per:parents”. Another example can be found in Table 4 (fourth instance), where the Mask model confused “org:top members/employees” with “per:employee of”. As we highlighted in our qualitative analysis in the main paper, Mask also makes mistakes because it fails to take into account the semantic types of the entities.

Encoder Predictions for the [MASK] token

In Table 7, we present examples of the tokens which are predicted by the language model for the [MASK] position of the appended relation prompts. Specifically, we use our pre-trained $[H, T] + Mask$ encoder that was initialised using RoBERTa-large. We consider the top five tokens directly predicted for the [MASK] positions by the encoder. We compare these predictions with the predictions from our full $[H, T] + Mask$ model, which comes with a classifier on top of the encoder. As can be seen, the token predictions do not adequately capture the relationships that are expressed in the given sentences. This is to be expected, since the InfoNCE loss which is used during pre-training encourages the embeddings of sentences that express the same relationship to be similar, but these vectors are no longer aligned with the tokens from the encoder’s vocabulary. This illustrates the need for a classifier that maps the embeddings of the [MASK] token in the relation prompt to dataset-specific relation labels.

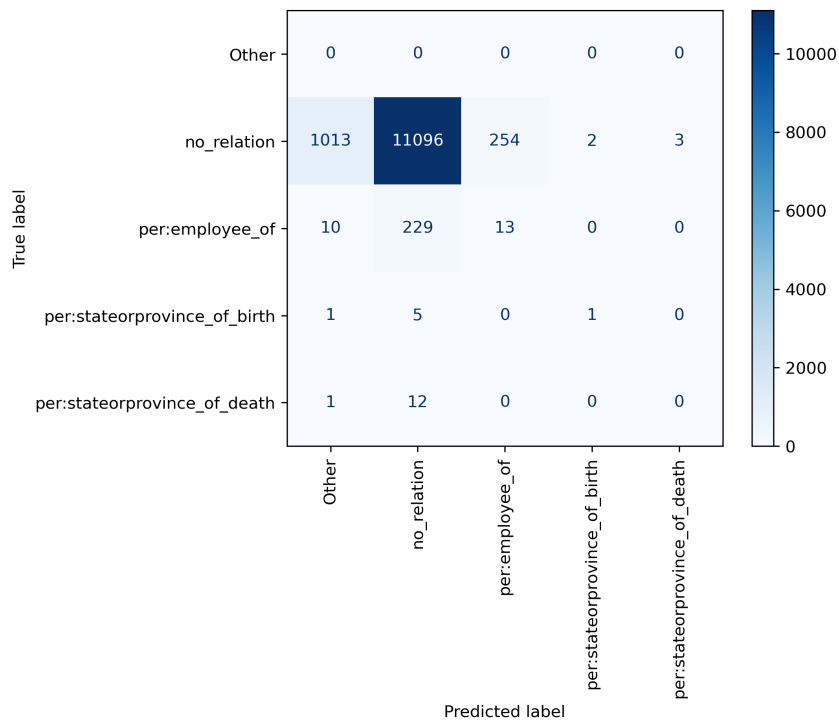


Figure 3: The confusion matrix for the Mask model on the TACREV test set.

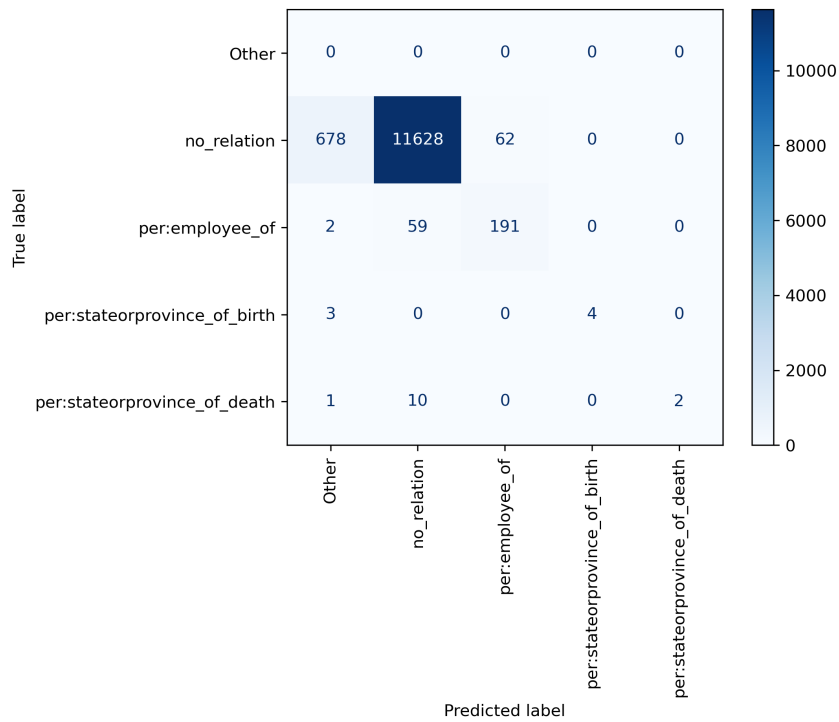


Figure 4: The confusion matrix for the [H,T] model on the TACREV test set.

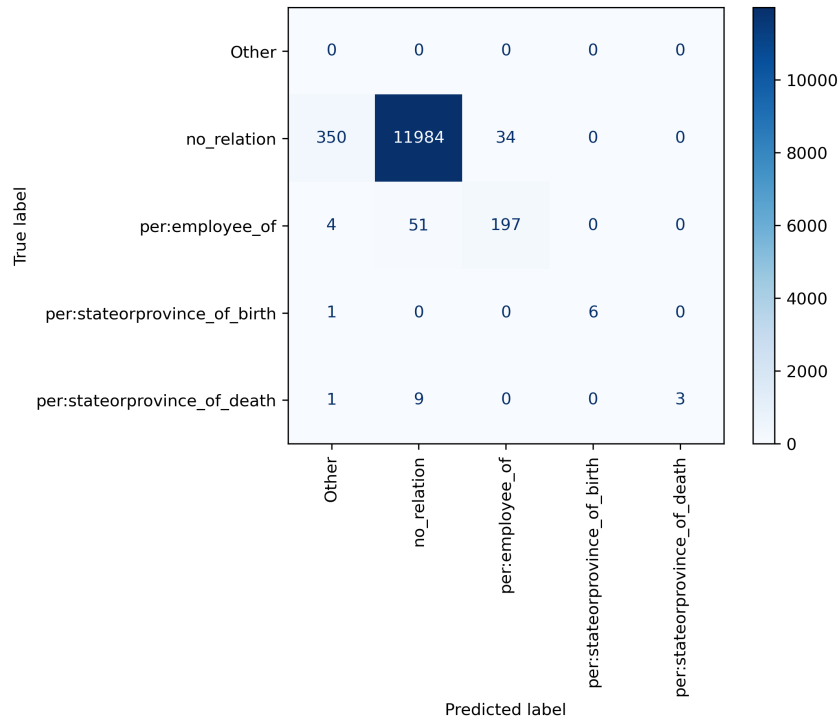


Figure 5: The confusion matrix for the [H,T]+ Mask model on the TACREV test set.

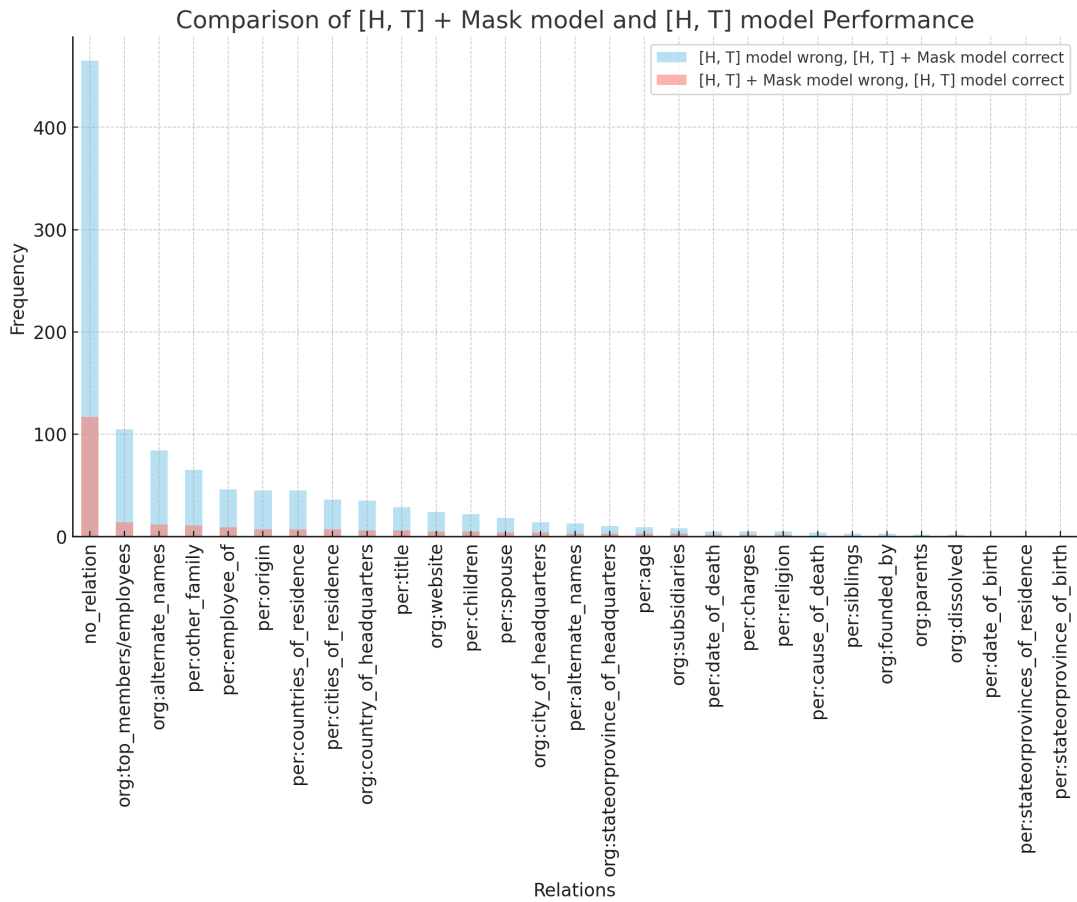


Figure 6: The comparison of the errors made by the [H,T] and [H,T]+Mask models on the TACREV test set.

Sentence	Label
“The current attempt to restore the commission was masterminded by a suspected mobster, Matteo Messina Denaro , who is among a handful of people vying to replace Provenzano,” police said.	Gold: per:title Mask: no relation [H,T]: no relation [H,T]+Mask: per:title
She ended up leaving Iraq under the threat of losing her job and returning home to Texas to seek medical and psychiatric treatment for post traumatic stress syndrome.	Gold: per:state or provinces of residence Mask: no relation [H,T]: no relation [H,T]+Mask: per:state or provinces of residence
Wen was tried with his wife, Zhou Xiaoya , and three former Chongqing police associates all of whom received jail sentences of up to 20 years.	Gold: per:spouse Mask: no relation [H,T]: no relation [H,T]+Mask: per:spouse
Knox’s father, Curt Knox, said his daughter looked “confident in what she wants to say.”	Gold: per:children Mask: no relation [H,T]: no relation [H,T]+Mask: per:children
Her mother, 60-year-old Claudie Mamane , tried to jump from the van while it was still moving and injured her arm.	Gold: per:parents Mask: no relation [H,T]: no relation [H,T]+Mask: per:parents
He is also survived by his parents and a sister, Karen Lange, of Washington, and a brother, Adam Lange , of St. Louis.	Gold: per:siblings Mask: per:parents [H,T]: no relation [H,T]+Mask: per:siblings
After more than 20 years of wearing the same long hairstyle, silver-tressed Ponce Kiah Marchelle Heloise Cruse Evans , known simply as Heloise , has a new ’do for the New Year.	Gold: per:alternate names Mask: no relation [H,T]: no relation [H,T]+Mask: per:alternate names

Table 5: Comparison of the TACREV test set predictions from the Mask, [H, T] and [H, T] + Mask models that were initialised using RoBERTa-large, focusing on examples where [H,T] incorrectly predicts “no relation”.

Sentence	Label
A man who shot and killed three women in a Pennsylvania health club, then himself, apparently blogged his rage-filled preparations –with the final chilling entry announcing the “big day” of the massacre.	Gold: per:state or province of death Mask: no relation [H,T]: per:state or provinces of residence [H,T]+Mask: per:state or province of death
Her brother-in-law, Wen , served as a top Chongqing police official for 16 years before taking over the city’s judiciary.	Gold: per:other family Mask: per:siblings [H,T]: per:siblings [H,T]+Mask: per:other family
Robert Holden, deputy director at the National Congress of American Indians , said “ the Washington DC-based group is hopeful that the use of secured cards could be expanded to allow tribal members to travel abroad.”	Gold: org:state or province of headquarters Mask: no relation [H,T]: per:state or provinces of residence [H,T]+Mask: per:state or provinces of residence
Countrywide Financial Corp co-founder Angelo Mozilo retired amidst scandal and investigation.	Gold: org:founded by Mask: no relation [H,T]: org:top members/employees [H,T]+Mask: org:founded by
Lt. Assaf Ramon , the son of Israel’s first astronaut, Col. Ilan Ramon , who died in the space shuttle Columbia disaster in 2003, was killed Sunday when an F16-A plane he was piloting crashed in the hills south of Hebron in the West Bank.	Gold: per:children Mask: per:parents [H,T]: per:siblings [H,T]+Mask: per:children

Table 6: Comparison of the TACREV test set predictions from the Mask, [H, T] and [H, T] + Mask models that were initialised using RoBERTa-large, focusing on examples where [H,T] confuses the ground truth with a different relation between entities of the same type.

Sentence	Label
Asia Bibi was sentenced to hang in Pakistan’s central province of Punjab earlier this month after being accused of insulting the Prophet Mohammed in 2009.	Gold: no relation Token predictions: strained, complicated, tense, fraught, complex. [H,T]+Mask: no relation
Benjamin Chertoff is the Editor in Chief of Popular Mechanics magazine as well as the cousin of the Director of Homeland Security , Michael Chertoff.	Gold: no relation Token predictions: unclear, unknown, complicated, complex, clear. [H,T]+Mask: no relation
“The current attempt to restore the commission was masterminded by a suspected mobster, Matteo Messina Denaro , who is among a handful of people vying to replace Provenzano,” police said.	Gold: per:title Token predictions: unclear, unknown, known, murky, complex. [H,T]+Mask: per:title
A man who shot and killed three women in a Pennsylvania health club, then himself, apparently blogged his rage-filled preparations –with the final chilling entry announcing the “big day” of the massacre.	Gold: per:state or province of death Token predictions: unclear, unknown, clear, murky, chilling. [H,T]+Mask: per:state or province of death
Her brother-in-law, Wen , served as a top Chongqing police official for 16 years before taking over the city’s judiciary.	Gold: per:other family Token predictions: unclear, complicated, unknown, strained, complex. [H,T]+Mask: per:other family
Robert Holden, deputy director at the National Congress of American Indians , said “ the Washington DC-based group is hopeful that the use of secured cards could be expanded to allow tribal members to travel abroad.”	Gold: org:state or province of headquarters Token predictions: unclear, confidential, unknown, complex, complicated. [H,T]+Mask: per:state or provinces of residence

Table 7: Comparison of the tokens predicted for the [MASK] position when using the pre-trained [H, T] + Mask encoder that was initialised using RoBERTa-large. Examples were selected from the TACREV test set. The token predictions are arranged in decreasing order of confidence (score).

C Ablations and Additional Experiments

Performance of the [CLS] token The intuition behind the Mask strategy is that the embedding of the [MASK] token captures the relational context. Another possible approach is to use the [CLS] token for this purpose. Table 8 analyses a number of variants that are based on this idea. Specifically, we consider the [CLS] token on its own, as well as variants of the [H,T,Mask] and [H,T]+Mask strategies where the role of the [MASK] token is replaced by the [CLS] token. These latter two strategies are referred to as [H,T,CLS] and [H,T]+CLS. For this analysis, we have used TACREV with standard pre-training. Overall, we can draw the same conclusions as for the variants with the [MASK] token, in particular when it comes to comparing [H,T,CLS] with [H,T]+CLS. However, the [CLS] variants yield results which are somewhat worse than the counterparts based on [MASK] (with the exception of the case where [CLS] is used on its own). This justifies the use of a prompt with [MASK].

Impact of Masking Approaches Peng et al. (2020) highlighted the importance of masking entity spans during pre-training, with some probability, to prevent the model from relying too much on the entity names themselves. Indeed, models

which rely too much on the entity names are prone to learning shortcuts which hamper generalisation, a problem which is sometimes referred to as entity bias (Wang et al., 2022b). Inspired by these works, we further investigate the impact of different masking strategies during pre-training. Specifically, we look at the following four scenarios in Table 9. First, in the strategy labelled *No masking*, we do not mask any tokens in the input corpus. Second, in the case of *Mask entity spans*, for each entity, we mask the entire entity span with 15% probability. For the variant labelled *Mask entity span heads*, we merely mask out the syntactic heads of entities, for 15% of the entity spans. We find the head word using the SpaCy dependency parser⁶. This is motivated by the idea that the syntactic head is most likely to reveal the entity type. Masking this head was found to be beneficial when pre-training entity encoders for this reason (Mtumbuka and Schockaert, 2023). Finally, in the case of *Random tokens*, we randomly mask 15% of the tokens in the training corpus. These tokens include both tokens from entity spans and non-entity tokens. This is the strategy that we have used for the main experiments. We can see that masking random tokens gives us the best results.

Comparison of Language Models In Table 10, we compare the performance of roberta-large, which we have been using for our main experiments, with bert-base-uncased⁷, bert-large-uncased and albert-xxlarge-v1⁸. Unsurprisingly, we can see that bert-large-uncased outperforms bert-base-uncased. Furthermore, across all datasets, roberta-large outperforms bert-large-uncased, but the best results are obtained by albert-xxlarge-v1. This is consistent with the findings from Zhong and Chen (2021). The main advantage of albert-xxlarge-v1 is that it uses 4096-dimensional embeddings, compared to 1024-dimensional embeddings for bert-large-uncased and roberta-large, despite being smaller than the latter two models (due to parameter sharing across layers). As can be seen, for all language models, the main conclusions remain consistent with those reported in the main paper. For instance, we consistently

⁶<https://spacy.io/api/dependencyparser>

⁷https://huggingface.co/docs/transformers/model_doc/bert

⁸https://huggingface.co/docs/transformers/model_doc/albert

find that the Mask strategy in isolation yields the lowest performance, that [H,T,Mask] improves on [H,T] and that [H,T]+Mask achieves the best results. This further supports our main findings about the complementarity of the [H,T] and Mask representations.

Few-shot Relation Classification To complement our main results, we have carried out an evaluation in the few-shot setting, where only a few training examples per relation are available. We employ three commonly-used few-shot learning settings on the family of TACRED datasets: 4 training examples per relation (representing approximately 1% of the full TACRED training set), 16 training examples per relation (approximately 5%), and 32 examples per relation (approximately 10%) (Sainz et al., 2021; Lu et al., 2022). We compare our models with the methods from Sainz et al. (2021) and Lu et al. (2022), which were specifically designed for the few-shot setting. In particular, Sainz et al. (2021) relied on pre-trained NLI models to solve relation extraction in the few-shot setting. Note that their approach relies on manually constructed verbalisations of the relations, which essentially provides an additional supervision signal. They reported results for NLI models based on RoBERTa-large (shown as $NLI_{RoBERTa}$) on DeBERTa-v2xxlarge (shown as $NLI_{DeBERTa}$). Lu et al. (2022) treat few-shot relation extraction as a summarisation task, relying on a pretrained PEGASUS-large abstractive summarisation model (shown as $SURE_{PEGASUS}$). They also rely on manually constructed verbalisations.

As can be seen in Table 11, the [H,T]+Mask model consistently outperforms $SURE_{PEGASUS}$. Furthermore, [H,T]+Mask outperforms $NLI_{RoBERTa}$ in all cases apart from the 1% setting, and $NLI_{DeBERTa}$ for the 10% and 100% configurations. This is remarkable, given that these baselines were specifically designed for the few-shot setting, rely on extensive pre-training, and in the case of $NLI_{DeBERTa}$ and $SURE_{PEGASUS}$ rely on much larger LMs. When it comes to the relative performance of the different variants that are considered in this paper, our overall findings are similar as for the main experiments. For instance, [H,T]+Mask and Encore+Mask achieve the best results, the performance of EnCore+Mask is again similar to that of [H,T]+Mask, and pre-training on Gigaword consistently improves the results. The performance of Mask and [H,T] is particularly

TACREV	
Mask	22.9
[H, T, Mask]	73.8
[H, T] + Mask	84.6
CLS	25.4
[H, T, CLS]	71.6
[H, T] + CLS	79.2

Table 8: Evaluation of strategies using the [CLS] token on TACREV. For this analysis, we have used RoBERTa-large with the standard pre-training strategy.

TACREV	
No masking	58.7
Mask entity spans	73.8
Mask entity span heads	79.4
Mask random tokens	84.6

Table 9: Evaluation of different masking strategies. For this analysis, we have used the [H, T] + Mask approach with RoBERTa-large and standard pre-training.

poor in the few-shot setting, and combining these two types of representations has a very big impact here. For instance, in the setting with 1% of the training data with standard pre-training, the performance increases from 19.7% for [H,T] to 48.4% for [H,T,Mask] and 53.0% for [H,T]+Mask.

			BERT-base	BERT-large	RoBERTa-large	ALBERT-xxl
Mask	TACRED	F1	20.9	22.2	23.3	25.1
	TACREV	F1	20.6	21.7	22.9	24.8
	Re-TACRED	F1	21.3	22.1	23.2	25.4
	Wiki-WD	F1	44.9	46.2	48.9	51.3
	NYT-FB	P@10	48.2	49.5	52.3	53.7
	NYT-FB	P@30	47.6	48.2	49.7	51.2
[H, T]	TACRED	F1	58.6	59.3	61.4	62.9
	TACREV	F1	59.1	60.7	63.6	64.3
	Re-TACRED	F1	67.4	69.1	72.3	73.5
	Wiki-WD	F1	72.4	73.1	75.2	76.9
	NYT-FB	P@10	75.2	76.7	79.6	80.7
	NYT-FB	P@30	73.9	75.2	78.5	79.3
[H, T, Mask]	TACRED	F1	70.1	71.6	73.0	74.7
	TACREV	F1	70.8	71.1	73.8	75.2
	Re-TACRED	F1	76.3	78.4	81.9	84.7
	Wiki-WD	F1	76.2	78.7	81.7	82.4
	NYT-FB	P@10	80.5	84.3	88.4	90.2
	NYT-FB	P@30	79.9	82.7	86.1	89.4
[H, T] + Mask	TACRED	F1	74.8	75.1	78.5	78.9
	TACREV	F1	75.6	77.3	84.6	83.9
	Re-TACRED	F1	86.4	87.9	91.9	92.3
	Wiki-WD	F1	79.7	84.9	89.6	89.8
	NYT-FB	P@10	93.9	94.1	94.7	94.9
	NYT-FB	P@30	93.4	92.5	93.1	93.8

Table 10: Comparison of different language models, using different relation embedding strategies under standard pre-training.

	F1			
	1%	5%	10%	100%
<i>Baselines</i>				
SpanBERT [†]	0.0	28.8	1.6	70.8
RoBERTa [†]	7.7	41.8	55.1	71.3
LUKE [†]	17.0	51.6	60.6	72.0
NLI [†] _{RoBERTa}	56.1	64.1	67.8	71.0
NLI [†] _{DeBERTa}	63.7	69.0	67.9	73.9
SURE [*] _{PEGASUS}	52.0	64.9	70.7	75.1
<i>Standard pre-training</i>				
Mask	8.3	11.6	15.0	23.3
[H, T]	19.7	27.5	33.9	61.4
[H, T, Mask]	48.4	54.6	61.3	73.0
[H, T] + Mask	53.0	65.9	71.7	78.5
EnCore + Mask	52.9	65.8	71.6	78.9
<i>Gigaword pre-training</i>				
Mask	9.9	13.1	18.3	24.6
[H, T]	21.1	28.7	34.3	63.2
[H, T, Mask]	49.2	55.1	62.1	74.4
[H, T] + Mask	53.5	66.1	71.9	78.5
EnCore + Mask	53.4	66.2	72.1	79.1

Table 11: Few-shot scenario results on TACRED with 1%, 5%, 10% and 100% of training data. [H, T] + Mask was initialised using RoBERTa-large and standard pre-training on the reduced training sets are used for these experiments. The results for models marked with [†] were taken from Sainz et al. (2021), whereas those marked with * were taken from Lu et al. (2022).