ACME-RE: Adaptive Context Memory and Evidence-guided Relation Extraction for Document-Level Relation Extraction

Anonymous ACL submission

Abstract

Document-level relation extraction (DocRE) is the task of identifying all relations between each entity pair in a document. Current methods still have room for improvement in handling implicit relationships, which are relations not explicitly stated in the text but can be inferred from the context. To address this limitation, we introduce the concept of context informativeness for entity pairs and propose ACME-RE (Adaptive Contextual Memory-Enhanced Relation Extraction), a novel framework for document-level relation extraction (DocRE). By introducing Evidence-guided context informativeness and an adaptive category memory module, ACME-RE significantly improves the performance of implicit relationship extraction. Experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance on the Re-DocRED dataset. This research provides a more comprehensive solution for document-level relation extraction and offers valuable insights for future studies.

Willi Schneider (skeleton racer)

[1] Wilfried "Willi" Schneider (born 13 March 1963 in Mediaş, Transylvania) is a German skeleton racer who competed from 1992 to 2002.[2] … [5] After retiring from competition, Schneider became a coach, leading the ***Canadian*** skeleton team to three medals at the 2006 Winter Olympics in Turin (a gold for ***Duff Gibson***, a silver for Jeff Pain, and a bronze for Melissa Hollingsworth), and coaching Jon Montgomery to victory in the 2010 Winter Olympics in Vancouver, British Columbia, ***Canada***.[6] In July 2012, Schneider agreed to a two-year contract to coach the *Russian* skeleton team.

Subject: ***Duff Gibson***	**Relation:** ***nationality***
Object: ***Canada***	**Evidence:** **5**

Figure 1: Example document and relation triple from DocRED, where sentences are numbered with [i]. Evidence sentences for this triple are shown in black, while non-evidence sentences are in grey. Subject and object mentions are shown in bold italics, and other entity mentions are underlined. Tokens with a red background indicate parts that require more attention, with the shade of red representing the level of attention.

1 Introduction

Relation extraction is a crucial task in natural language processing that aims to categorize relationships between two specified entities into predefined classes. While sentence-level relation extraction (RE) has made significant progress (Peng et al., 2017; Verga et al., 2018; Yao et al., 2019), document-level relation extraction (DocRE) faces substantial challenges, particularly with implicit relations that are not explicitly stated in the text. These implicit relations are vital for applications such as knowledge graph construction and question answering enhancement. For instance, as shown in Figure 1, a DocRE system must infer the nationality relationship between Duff Gibson and Canada, even when it's not directly stated.

A novel observation we make is that for such implicit information, even when inferred from the source text, the context containing this triple carries minimal information about the relationship (e.g., nationality) since it's not the primary focus of the document's narrative. This observation leads us to propose leveraging other instances of the same relation type, particularly those with richer contextual information, to enhance the representation of triples with limited information.

Existing DocRE approaches primarily focus on explicit relations, employing customized loss functions and document-level processing to address label imbalance and complexity issues (Zhou et al., 2021; Tan et al., 2022a). However, these methods struggle when handling implicit relations. Few approaches effectively address relationships requiring deep contextual understanding, and none adequately handle the varying amounts of contextual information across different triples. While memory-augmented models like TTM-RE (Gao et al., 2024) enhance context by leveraging previously encountered entities and scenarios, we argue that memorizing entities introduces unnecessary additional parameters in large-scale document-level relation extraction tasks. Instead, we aim to learn an augmentable context vector for each relation category to improve existing relation prediction performance.

To address these challenges, we propose ACME-RE (Adaptive Contextual Memory-Enhanced Relation Extraction), a novel framework that dynamically adapts to varying levels of contextual information. Rather than memorizing specific entities, ACME-RE employs a memory module that maintains and updates category-specific contextual patterns, integrating both explicit evidence sentences and implicit contextual cues. The evidence sentences are manually annotated parts of the original text that support the relationship in the triplet. Another related concept is the context of an entity pair, which is obtained by applying the attention of the entity pair over all sentences and tokens in the document, multiplied by the full document embedding. This adaptive approach enables the model to effectively handle cases where direct evidence is insufficient by leveraging learned patterns from information-rich instances of the same relation type.

Our contributions are: (1) The ACME-RE framework, the first designed to address contextual variability across triplets, enabling deep contextual understanding and inference. (2) State-of-the-art performance, achieved with minimal additional parameters, on benchmark datasets using both gold and distantly supervised data.

2 Preliminary

2.1 Problem Definition

Given a document D consisting of sentences $X_D = \{x_i\}_{i=1}^{|X_D|}$ and entities $E_D = \{e_i\}_{i=1}^{|E_D|}$. Each entity $e \in E_D$ appears at least once in D, with its mentions denoted as $M_e = \{m_i\}_{i=1}^{|M_e|}$. As each pair of entities (e_s, e_o) can have multiple relations, the goal of document-level relation extraction (DocRE) is to predict a set of relations $R_{s,o} \subset R$, where R is a set of predefined relations. Given the N entities in D, the model needs to consider up to $R \times N \times (N-1)$ possible relations.

2.2 DREEAM

DREEAM (Ma et al., 2023) enhances the ATLOP model by integrating evidence information into the attention mechanism (details can be found in Appendix B). It supervises the attention module to focus on evidence sentences while reducing attention to irrelevant text. Since the distantly supervised dataset lacks evidence annotations, the method proposes a distillation-based three-stage training framework. First, it utilizes human-annotated data for supervision and uses the teacher model as an evidence distribution predictor to predict the evidence distribution of the distantly supervised data. This distribution is then used as a learning signal to train the student model. Finally, the student model is fine-tuned on the human-annotated dataset to obtain the final model. For specific details, please refer to Appendix C.

DREEAM effectively utilizes both distantly supervised data and evidence annotation information. However, in the first stage of training, it aligns the attention distribution predicted by the model for entity pairs with the human-annotated evidence distribution. Specifically, the evidence sentences are assigned a total attention weight of 1, while non-evidence sentences receive attention weights that are infinitely close to zero. This approach of enhancing context based on evidence annotation information is somewhat coarse when addressing the issue of imbalanced triplet information quantity, as shown in Figure 1. Therefore, we aim to enhance the context of entity pairs more precisely based on contextual information quantity.

2.3 TTM-RE

TTM-RE is a memory-augmented framework for document-level relation extraction (DocRE) that integrates Token Turing Machine (Ryoo et al., 2023) memory modules and a noise-suppressing loss function (SSR-PU). While it addresses some limitations of existing methods in utilizing large-scale, noisy training data through memory-enhanced representations and robust handling of false negatives, its entity-centric memory mechanism reveals several inherent limitations.

Specifically, TTM-RE enhances inference by incorporating extra-document information about entities and iteratively updates entity combinations in its memory module to store the most representative pairs. This approach essentially memorizes entity type information and entity-specific patterns related to predefined relations. Although this mechanism can partially alleviate the insufficient information problem illustrated in Figure 1, it suffers from two critical drawbacks: (1) the category-relevant information is scattered across individual entities, making it indirect and fragmented, and (2) the entity-based memory unit introduces a substantial number of additional parameters.

To address these limitations, we propose an adaptive contextual memory mechanism that operates at the category level rather than the entity level. This novel approach not only significantly reduces the number of additional parameters but also learns more abstract and generalizable category vectors. By directly modeling relation categories, our method captures more comprehensive and coherent patterns, leading to more efficient and effective relation extraction.

3 Proposed Method: ACME-RE

We propose ACME-RE, an adaptive context memory and evidence-guided document-level relation extraction method. An illustration of the overall framework of ACME-RE is shown in Figure 2

3.1 Evidence-guided context informativeness

In document-level relation extraction, a document typically contains multiple entities, leading to an exponential increase in the number of entity pairs. While each document primarily describes key events, the semantic information is concentrated on a limited set of predefined relations relevant to these events. In contrast, many implicit relations beyond the main event are more challenging to identify. Moreover, document-level relation extraction requires the simultaneous identification of all possible relations among entity pairs, making it even harder to predict relations involving non-primary entities.

To address this challenge, we introduce the concept of context informativeness for entity pairs. Given a triple (h, r, t), the Semantic Information Quantity f the entity pair context (hereafter referred to as "information quantity") measures the extent to which the context expresses the relation r between the head entity h and the tail entity t. A higher information quantity indicates that the relation r between h and t can be inferred more easily from the context, while a lower information quantity suggests the opposite.

The information quantity is influenced by factors such as the relevance, clarity, and specificity of evidence sentences linking h and t. If the context explicitly mentions r, provides detailed descriptions, or establishes a strong semantic connection between h and t, the information quantity is typically high. Conversely, if the context is vague, ambiguous, or lacks sufficient relational cues, the information quantity is lower.

The specific implementation is as follows. Let the input sequence be represented as $X \in R^{bs \times hw \times c}$, where bs is the batch size, hw is the number of tokens (including memory and context tokens), and c is the feature dimension. Each token's information quantity is computed as:

$$I(x_i) = \text{MLP}(\text{LayerNorm}(x_i)) \tag{1}$$

where x_i is the i-th token, $\text{LayerNorm}(\cdot)$ applies layer normalization to standardize the input, and $\text{MLP}(\cdot)$ is a multi-layer perceptron mapping the normalized input to a scalar information quantity.

We quantify the information quantity of an entity pair's context to assess its effectiveness as evidence. Additionally, by incorporating category-based contextual memory, we enhance the original context, improving the accuracy of relation inference.

In document-level relation extraction, DREEAM leverages evidence sentence distributions to guide entity pair context modeling. However, sentence-level annotations are not entirely accurate—some tokens within evidence sentences are uninformative, while useful tokens may exist outside these sentences. Our information quantity measure mitigates this limitation by enabling relevant tokens outside evidence sentences to contribute, thus extracting a more effective entity pair context.

3.2 Adaptive category memory module

Earlier relation extraction methods generally obtained relation embeddings by concatenating the head and tail entity embeddings. However, this approach primarily captures entity type information rather than specific relational information, leading to false positives when relations involve entities of the same type.

ATLOP (Zhou et al., 2021) and DREEAM (Ma et al., 2023) partially addressed this limitation by enhancing the contextual representations of entity pairs. Meanwhile, (Mtumbuka and Schockaert, 2023) proposed a sentence-level relation extraction method based on the [MASK] token, which learns a [MASK] vector to supplement entity pair type information.

However, unlike sentence-level extraction, where the number of entity pairs is relatively limited, document-level relation extraction faces an exponential increase in entity pairs, making it in-

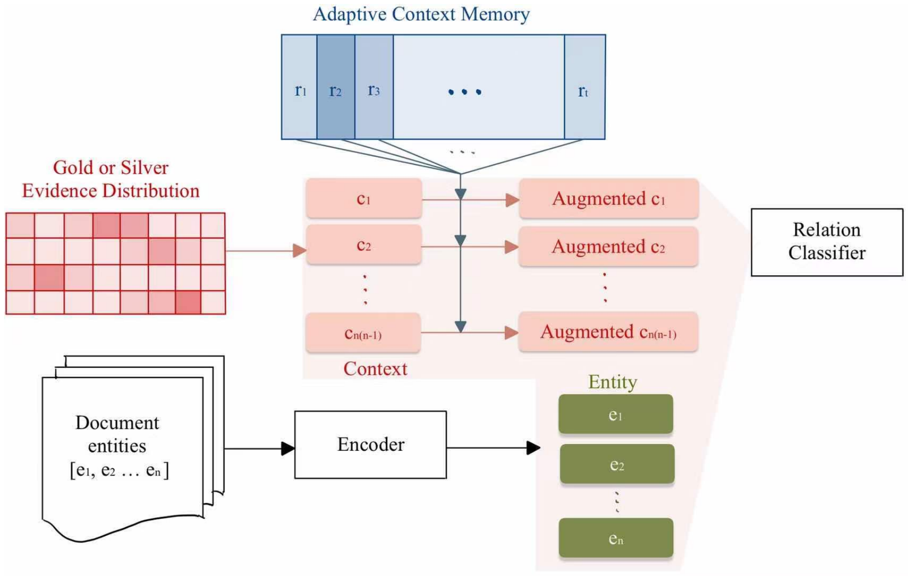

Figure 2: Overall framework of ACME-RE. The number of memory vectors corresponds to the number of predefined categories(e.g., $r_1, r_2, \ldots, r_t$).The depth of the memory vector's color represents the amount of information.

feasible to introduce a similar [MASK] mechanism directly. This necessitates an efficient strategy to incorporate category vectors without introducing excessive computational overhead.

Inspired by TTM (Gao et al., 2024), one possible solution is to use memory vectors (Ryoo et al., 2023), dynamically selecting relevant category memory vectors based on contextual evidence cues and all category memory vectors. This ensures scalability while preserving rich relational information. Specifically, category memory vectors encode prototypical representations of predefined relation types, which can be leveraged to enhance relation embeddings and mitigate the limitations of both entity concatenation-based and evidence-context-based approaches.

Building upon contextual information quantification, the specific implementation of category memory vectors is as follows. First, the relative contribution of each token is obtained by applying a softmax function:

$$S_i = \frac{\exp(I(x_i))}{\sum_{j=1}^{hw} \exp(I(x_j))} \tag{2}$$

where S_i represents the normalized contribution of token i to the aggregated representation.

The enhanced contextual representation is computed as:

$$Z = \sum_{i=1}^{hw} S_i \cdot x_i \tag{3}$$

where $Z \in R^{bs \times n_{token} \times c}$ is the transformed representation with enhanced contextual information. Notably, this ensures that tokens with higher information content contribute more significantly to the final representation. Consequently, when a category contextual vector is frequently evaluated as low-information, it retains more information in the memory, and vice versa.

Subsequently, the enhanced context and evidence context are averaged to obtain the final context representation (experimental results show that averaging slightly outperforms concatenation), which, together with entity pair embeddings, serves as the basis for relation prediction to reduce the number of parameters to enable more efficient learning, during which we adopt Group Bilinear Classification, where augmented entity representations are split into k parts with dimension (d/k):

$$p(r|e'_h, e'_t) = \sigma\left(\sum_{i=1}^{k} e'^{(i)}_h B_i e'^{(i)}_t\right),$$

where $B_i \in R^{d/k \times d/k}$ are learnable bilinear parameters. This reduces the parameter count from d^2 to d^2/k, improving efficiency.

Additionally, to effectively utilize distant supervision data, this study adopts a three-stage training framework inspired by DREEAM (Ma et al., 2023). The update of category memory vectors follows a mechanism similar to [MASK], where they are randomly initialized and automatically updated via backpropagation during the gold data training stage. However, category memory vectors remain frozen during the subsequent distant supervision and fine-tuning stages.

Noise-Robust Loss Function (SSR-PU) To mitigate false negatives in distantly supervised data (Gao et al., 2023), we employs a Self-Supervised Robust Positive-Unlabeled (SSR-PU) loss, as done in TTM-RE:

Firstly, Traditional PU learning assumes that the overall data distribution aligns with the unlabeled data distribution, which may not hold in our case (Charoenphakdee and Sugiyama, 2019). To address this issue, it is necessary to consider PU learning under prior shift (Wang et al., 2022; Du Plessis et al., 2015, 2014).

For each class, let the original prior be $\pi_i = p(y_i = +1)$, and define the labeled prior as $\pi_{\text{labeled},i} = p(s_i = +1)$, where $s_i = +1$ or $s_i = -1$ indicates whether the i-th class is labeled or unlabeled, respectively. Then, the probability of an unlabeled sample being positive is:

$$\pi_{u,i} = p(y_i = 1 | s_i = -1) = \frac{\pi_i - \pi_{\text{labeled},i}}{1 - \pi_{\text{labeled},i}},$$

The non-negative risk estimator under class prior shift of training data is obtained as follows (Wang et al., 2022; Kiryo et al., 2017):

$$\hat{R}_{S-PU}(f) = \sum_{i=1}^{K} \Bigg(\frac{\pi_i}{n_{P_i}} \sum_{j=1}^{n_{P_i}} \ell(f_i(x_j^{P_i}), +1) + \max \Bigg(0, \Bigg[\frac{1}{n_{U_i}} \frac{1-\pi_i}{1-\pi_{u,i}} \sum_{j=1}^{n_{U_i}} \ell(f_i(x_j^{U_i}), -1) - \frac{1}{n_{P_i}} \frac{\pi_{u,i} - \pi_{u,i}\pi_i}{1-\pi_{u,i}} \sum_{j=1}^{n_{P_i}} \ell(f_i(x_j^{P_i}), -1) \Bigg] \Bigg) \Bigg).$$

where $\pi_i = p(y_i = +1)$ denotes the probability of positive prior for relation class i. n_{P_i} and n_{U_i} are the numbers of positive and unlabeled samples of class i, respectively. ℓ is a convex loss function, and $f_i(\cdot)$ is a score function that predicts class i. $x_j^{P_i}$ and $x_j^{U_i}$ denote that the j-th sample of class i is positive and unlabeled as class i, respectively. This formulation ensures robust learning under noisy, unlabeled data. For more details, we refer the readers to the original paper (Wang et al., 2022; Tang et al., 2022)

Table 1: Statistics of the DocRED dataset and Re-DocRED dataset. In total, there are 96 relations. The distantly supervised dataset is the same as in DocRED and is created with no human supervision.

Statistics	Distant	DocRed	Re-D
# Docs	101,873	5,053	4053
Avg. # Entities	19.3	19.5	19.4
Avg. # Labeled Triples	14.8	12.5	29.7
Avg. # Sentences	8.1	8.0	7.9

4 Experiments

4.1 Setting

Datasets To evaluate our approach, we primarily use the DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b) datasets. DocRED includes manually annotated data and distant supervision data generated by aligning Wikipedia with Wikidata (Vrandečić and Krötzsch, 2014). Re-DocRED addresses the incompleteness and logical inconsistencies present in the original DocRED dataset and corrects coreference errors. Table 1 shows the amount of training data available for all data splits as well as the average number of entities.

Configuration We implement ACME-RE based on Hugging Face's Transformers (Wolf, 2020).Following previous work, we evaluate the performance of DREEAM using RoBERTa-large (Liu, 2019) as the PLM encoder. The parameter for balancing ER loss with RE loss is set to 0.05 when training both the teacher and the student model, chosen based on a grid search from 0.05, 0.1, 0.2, 0.3. We train and evaluate ACME-RE on a single NVIDIA A800 80GB GPU. Details about hyper-parameters and running time will be provided in Appendix A.

Evaluation For evaluation, we adopt official evaluation metrics of DocRED (Yao et al., 2019): Ign F1 and F1 for RE. Ign F1 is measured by removing relations present in the annotated training set from the development and test sets. We train our system five times, initialized with different random seeds, and report the average scores and standard error of these runs.

4.2 Main Results

Table 2 lists the performance of the proposed and existing methods. We select the best-performing model on the development set to make predictions on the test set.

Table 2: Evaluation results on test set of Re-DocRED, with best scores bolded. The scores of existing methods are borrowed from corresponding papers.

Method	Ign F1	F1
(a) without Distantly-Supervised Data		
ATLOP (Zhou et al., 2021)	76.82	77.56
DocuNet (Zhang et al., 2021)	77.26	77.87
KD-DocRE (Tan et al., 2022a)	77.60	78.28
TTM-RE (Gao et al., 2024)	78.20	79.95
DREEAM-teacher (Ma et al., 2023)	79.66	80.73
ACME-RE(Ours)	**80.25**$_{\pm0.24}$	**81.21**$_{\pm0.19}$
(b) with Distantly-Supervised Data		
ATLOP (Zhou et al., 2021)	78.52	79.46
DocuNet (Zhang et al., 2021)	79.41	80.37
KD-DocRE (Tan et al., 2022a)	80.32	81.04
TTM-RE (Gao et al., 2024)	83.11	84.01
DREEAM-student (Ma et al., 2023)	80.39	81.44
ACME-RE(Ours)	**83.67**$_{\pm0.20}$	**84.61**$_{\pm0.17}$

Results on Re-DocRED. From Table 2, our proposed method achieves state-of-the-art performance, outperforming all existing approaches across multiple evaluation metrics. Specifically, compared to DREEAM (the best method using evidence), our model improves the F1 score by X and Y under human-annotated data and combined data settings, respectively, with IgnF1 gains of X_1 and Y_1. Similarly, compared to TTM-RE (the best method without evidence), our model achieves F1 improvements of X and Y, and IgnF1 improvements of X_1 and Y_1 under the same settings. These results highlight the effectiveness of ACME-RE's evidence utilization mechanism in capturing complex patterns and relationships within the data, aligning with prior work on entity-based and masked-prompt strategies (Genest et al., 2022; Zhong and Chen, 2020).

Furthermore, consistent performance gains on both development and test sets demonstrate the robustness and generalizability of our approach. Notably, while TTM-RE's best results were achieved with a memory size of 200 (with no reports on larger sizes), our method requires less than half its memory capacity while delivering significantly better performance.

Results on DocRED. As shown in Table 3, while ACME-RE achieves comparable improvements on the DocRED validation set as on Re-DocRED, its performance on the DocRED test set is less satisfactory. We analyze this phenomenon from two perspectives:

- **Data Distribution**: ReDocRED's validation and test sets are equally split from DocRED's validation set, potentially creating a distribution mismatch with the original DocRED test set. While our method demonstrates strong performance on ReDocRED, it may encounter out-of-distribution challenges on the DocRED test set.
- **Methodology and Data Quality**: Our approach exhibits different behaviors on varying data qualities. While ReDocRED's improved annotation quality enables our memory module to learn precise relation prototypes, this precision becomes a limitation when encountering DocRED's test set where false negatives persist. This contrast explains our model's strong performance on ReDocRED's validation and test sets but relatively weaker results on DocRED's test set, as the precisely learned prototypes may not generalize well to noisier scenarios.

To address these challenges, future work could focus on enhancing the memory module's robustness through noise-aware training strategies. This would maintain our method's strength on high-quality data while improving its resilience to noisy instances.

4.3 Ablation Studies

This subsection investigates the effect of context-informativeness-guided memory and and evidence-guided training by ablation studies. All subsequent experiments adopt RoBerta-large as the PLM encoder.

Teacher Model Firstly, we explore how guiding attention through contextual information quantity can assist in training relation extraction (RE) on human-annotated data. To better detect the effect of adaptive category memory guided by information quantity, we compare it with the memory module of TTM-RE, which uses a single entity-context concatenation as a unit. The results, as shown in Table 4, indicate a significant decline in the RE performance of our system under this setup. To further analyze the importance of different components, we conduct ablation studies by training variants of our teacher model. Specifically, we create a variant without evidence extraction (ER) training and evaluate its performance on the Re-DocRED development set. When the contextual information quantity training is disabled, the model effectively degrades to a baseline model similar to TTM-RE

Table 3: Evaluation results on development and test sets of DocRED, with best scores bolded. The scores of existing methods are borrowed from corresponding papers. We group the methods first by whether they utilize the distantly-supervised data or not, then by whether they utilize evidence.

	Use Evidence	**Dev**		**Test**	
Method		**Ign F1**	**F1**	**Ign F1**	**F1**
(a) without Distantly-Supervised Data					
SSAN (Xu et al., 2021)	No	60.25	62.08	59.47	61.42
ATLOP (Zhou et al., 2021)	No	61.32	63.18	61.39	63.40
DocuNet (Zhang et al., 2021)	No	62.23	64.12	62.39	64.55
TTM-RE (Gao et al., 2024)	No	61.78	64.11	59.81	61.07
EIDER (Xie et al., 2021)	Yes	62.34	64.27	62.85	64.79
SAIS (Xiao et al., 2021)	Yes	62.23	65.17	**63.44**	**65.11**
DREEAM-teacher (Ma et al., 2023)	Yes	62.29	64.20	62.12	64.27
ACME-RE(Ours)	Yes	**62.95**$_{\pm 0.37}$	**65.32**$_{\pm 0.24}$	60.21$_{\pm 0.31}$	62.57$_{\pm 0.21}$
(b) with Distantly-Supervised Data					
SSAN (Xu et al., 2021)	No	63.76	65.69	63.78	65.92
KD-DocRE (Tan et al., 2022a)	No	65.27	67.12	65.24	67.28
TTM-RE (Gao et al., 2024)	No	67.99	70.00	65.11	66.98
DREEAM-student (Ma et al., 2023)	Yes	67.41	65.52	65.47	67.53
ACME-RE(Ours)	Yes	**70.38** $_{\pm 0.34}$	**72.17**$_{\pm 0.21}$	**66.48**$_{\pm 0.19}$	**67.88**$_{\pm 0.29}$

Table 4: Ablation studies evaluated on the Re-DocRED development set.

Setting	**Ign F1**	**F1**
(a) Teacher Model		
ACME-RE	**80.01**$_{\pm 0.24}$	**81.01**$_{\pm 0.19}$
w/o Adaptive memory	78.95$_{\pm 0.31}$	80.03$_{\pm 0.25}$
w/o ER training	78.15$_{\pm 0.23}$	79.86$_{\pm 0.14}$
(b) Student Model		
ACME-RE	**83.27**$_{\pm 0.21}$	**84.22**$_{\pm 0.14}$
w/o Adaptive memory	80.89$_{\pm 0.39}$	81.82$_{\pm 0.34}$
w/o ER training	80.14$_{\pm 0.18}$	81.03$_{\pm 0.17}$

Table 5: Effect of the size of the number of memory tokens available to be used in ACME-RE on the test dataset of Re-DocRED.

Mem. Size	**Ign F1**	**F1**
90	**79.79**$_{\pm 0.12}$	**80.86**$_{\pm 0.13}$
96	**80.25**$_{\pm 0.24}$	**81.21**$_{\pm 0.19}$
100	80.04$_{\pm 0.29}$	80.92$_{\pm 0.17}$
200	**79.69**$_{\pm 0.21}$	**80.67**$_{\pm 0.16}$
400	**79.46**$_{\pm 0.20}$	**80.42**$_{\pm 0.22}$
600	79.16$_{\pm 0.31}$	80.13$_{\pm 0.23}$

(Gao et al., 2024). As shown in Table 4, removing ER training also leads to a slight decline in model performance. These observations suggest that while the guidance of contextual information quantity plays a crucial role in entity-pair context learning, it cannot fully substitute the contribution of evidence information. This finding underscores the complementary nature of both components in achieving optimal RE performance.

Student Model Next, we investigate the student model, which undergoes a two-phase training process: initial training on distantly supervised data followed by fine-tuning on human-annotated data. Following the same experimental approach used for the teacher model, we conduct ablation studies to examine how adaptive category memory affects the model's performance at different training stages. The results, as shown in Table 4, reveal that ACME-RE experiences a more substantial performance degradation when adaptive category memory is removed compared to both the teacher model and the variant without evidence guidance, highlighting its crucial role in the student model's architecture. Notably, although ACME-RE may show suboptimal performance on noisy training data (like DocRED), the adaptive category memory vectors inherited from the teacher model, which is trained on high-quality annotated data, can effectively guide the training process. These pre-learned contextual category memory vectors serve as reliable guidance for filtering and utilizing information from distant supervision, while maintaining adaptability during subsequent fine-tuning on human-annotated data.

4.4 Memory Size

As shown in Table 5, the size of memory tokens significantly impacts ACME-RE's performance on the Re-DocRED test set. The model achieves its best performance with a memory size of 96, indicating that aligning the memory size with the number of predefined relations optimizes the model's ability to capture and utilize contextual information effectively.

Deviating from this optimal size, whether by increasing or decreasing it, leads to a decline in performance. Larger memory sizes, such as 600, introduce noise and redundancy, reducing the model's focus on relevant information. Conversely, smaller sizes, like 90, fail to fully capture the complexity of the data, resulting in suboptimal performance.

These findings underscore the importance of carefully tuning the memory size to balance contextual information capture and computational efficiency, ensuring optimal performance in relation extraction tasks.

5 Related Work

5.1 DocRE

Recent work has extended the scope of relation extraction task from sentence to document (Peng et al., 2017; Quirk and Poon, 2016). Current benchmarks include DocRED (Yao et al., 2019), re-DocRED (Tan et al., 2022b), CDR (Li et al., 2016) and GDA (Wu et al., 2019), among which, DocRED (Yao et al., 2019) and re-DocRED (Tan et al., 2022b) are notable for including both evidence annotations and distantly supervised data.

5.2 Transformer-based DocRE

Modern DocRE approaches are built upon Transformer-based pretrained language models, demonstrating superior performance in capturing long-distance dependencies (Yao et al., 2021; Zeng et al., 2020, 2021; Zhang et al., 2021). ATLOP (Zhou et al., 2021) established a strong baseline by introducing adaptive thresholding and localized context pooling for improving extraction accuracy. Building on ATLOP, various methods incorporate graph structures to enhance cross-sentence reasoning (Zhang et al., 2023). GAIN (Zeng et al., 2020) employs heterogeneous mention-level and entity-level graphs with path reasoning, SSAN (Xu et al., 2021) integrates entity structure dependencies, and TAG (Zhang et al., 2023) introduces latent graphs with hierarchical clustering. Recent advances focus on addressing key challenges through new loss functions (Tan et al., 2022a; Wang et al., 2022; Zhou and Lee, 2022; Wang et al., 2023) and memory mechanisms (Gao et al., 2024) to better handle class imbalance and leverage large-scale noisy data.

5.3 DocRE with Evidence

Evidence incorporation has evolved from heuristic approaches to neural methods. E2GRE (Huang et al., 2021a) pioneered heuristic evidence selection to enhance DocRE performance, an approach later adopted by (Huang et al., 2021b). Subsequent works (Xie et al., 2021; Xiao et al., 2021) developed neural classifiers for evidence retrieval. SAIS (Xiao et al., 2021) introduced hierarchical evidence retrieval, first identifying entity pair evidence sets before refining them for specific relations. Eider (Xie et al., 2021) guides attention weights using evidence sentences, while DREEAM (Ma et al., 2023) improves this through KL divergence-based alignment of attention and evidence distributions. Unlike previous approaches, our method combines evidence annotations with context informativeness without relying on heuristic rules or neural classifiers, providing more effective information for relation extraction.

6 Conclusion

In this paper, we introduce a novel framework called ACME-RE for document-level relation extraction (DocRE) that leverages adaptive contextual memory to address the challenge of extracting implicit relationships. The experimental results demonstrate that ACME-RE provides a robust and efficient solution for document-level relation extraction, particularly in scenarios requiring the inference of implicit relationships. Our findings pave the way for future research in memory-augmented techniques for information extraction tasks, offering insights into the balance between memory capacity and computational efficiency. We believe that it opens new avenues for exploring the integration of memory mechanisms in LLMs.

7 Limitations

Our work has several limitations that highlight avenues for future research. First, the method is sensitive to data quality, particularly false negatives, performing well with high-quality data but less effectively with noise, suggesting a need for more robust training. Second, the approach to modeling contextual information lacks granularity, ignoring factors like clarity and specificity, which could be addressed through a more refined decomposition of information components.

Ethical Statement

Based on the methodology employed in this study, we do not foresee any significant ethical concerns. All documents and models used in our research were obtained from open-source domains, ensuring transparency and accessibility. ACME-RE is trained exclusively on open-source document-level relation extraction data, which eliminates the risk of privacy leakage. Relation extraction is a well-established and widely studied task in natural language processing, with applications that are generally non-controversial.

The training process for ACME-RE required over 72 hours on NVIDIA A800 80GB GPUs, with the distantly supervised fine-tuning phase being particularly resource-intensive due to the dataset size. Derivatives of the data accessed for research purposes should not be used outside of research contexts. To promote reproducibility and further research, the code for ACME-RE will be released at a future date.

We believe that ACME-RE contributes positively to the field of document-level relation extraction and hope that its release will facilitate further advancements in memory-augmented models for natural language processing tasks.

References

Nontawat Charoenphakdee and Masashi Sugiyama. 2019. Positive-unlabeled classification under class prior shift and asymmetric error. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 271–279. SIAM.

Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. 2020. Hierarchical entity typing via multi-level learning to rank. *arXiv preprint arXiv:2004.02286*.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. 2015. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR.

Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. 2014. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27.

Chufan Gao, Xulin Fan, Jimeng Sun, and Xuan Wang. 2023. Promptre: Weakly-supervised document-level relation extraction via prompting-based data programming. *arXiv preprint arXiv:2310.09265*.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. Ttm-re: Memory-augmented document-level relation extraction. *arXiv preprint arXiv:2406.05906*.

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*, pages 561–571.

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021a. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 307–315.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021b. Three sentences are all you need: Local path enhanced document relation extraction. *arXiv preprint arXiv:2106.01793*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. *arXiv preprint arXiv:1904.02347*.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. Dreeam: Guiding attention with evidence for improving document-level relation extraction. *arXiv preprint arXiv:2302.08675*.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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.

Frank Mtumbuka and Steven Schockaert. 2023. Entity or relation embeddings? an analysis of encoding strategies for relation extraction. *arXiv preprint arXiv:2312.11062*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Chris Quirk and Hoifung Poon. 2016. Distant supervision for relation extraction beyond the sentence boundary. *arXiv preprint arXiv:1609.04873*.

Michael S Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. 2023. Token turing machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19070–19081.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. *arXiv preprint arXiv:2203.10900*.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred–addressing the false negative problem in relation extraction. *arXiv preprint arXiv:2205.12696*.

Zhenwei Tang, Shichao Pei, Zhao Zhang, Yongchun Zhu, Fuzhen Zhuang, Robert Hoehndorf, and Xiangliang Zhang. 2022. Positive-unlabeled learning with adversarial data augmentation for knowledge graph completion. *arXiv preprint arXiv:2205.00904*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. *arXiv preprint arXiv:1802.10569*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. 2023. Adaptive hinge balance loss for document-level relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3872–3878.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. *arXiv preprint arXiv:2210.08709*.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2021. Sais: Supervising and augmenting intermediate steps for document-level relation extraction. *arXiv preprint arXiv:2109.12093*.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2021. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. *arXiv preprint arXiv:2106.08657*.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14149–14157.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. *arXiv preprint arXiv:2106.13474*.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction. *arXiv preprint arXiv:2106.01709*.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. *arXiv preprint arXiv:2009.13752*.

Liang Zhang, Zijun Min, Jinsong Su, Pei Yu, Ante Wang, and Yidong Chen. 2023. Exploring effective inter-encoder semantic interaction for document-level relation extraction. In *IJCAI*, pages 5278–5286.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. *arXiv preprint arXiv:2106.03618*.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. *arXiv preprint arXiv:2205.00476*.

A Parameter Settings

We adopt AdamW as the optimizer (Loshchilov, 2017) and apply a linear warmup for the learning rate at the first 6% steps. Important hyperparameters are shown in Table

Table 6: Hyperparameters for Training and Fine-tuning

Hyperparam.	Teacher Train	Student Train	Student Finetune
# Epoch	40	2	40
lr for encoder	5e-5	3e-5	1e-6
lr for classifier	1e-4	1e-4	3e-6
max gradient norm	1.0	5.0	2.0

B ATLOP: Adaptive Thresholding and Localized Context Pooling

ATLOP (Zhou et al., 2021) is a Transformer-based model for document-level relation extraction (DocRE). It introduces entity-pair localized context embeddings and adaptive thresholding to effectively handle long documents and multi-entity interactions.

Text Encoding. Given a document D with tokens $T_D = \{t_i\}_{i=1}^{|T_D|}$, ATLOP inserts special tokens (*) at the boundaries of entity mentions and encodes the tokens using a pretrained language model (PLM) (Vaswani, 2017). The token embeddings $H \in R^{|T_D| \times d}$ and cross-token dependencies $A \in R^{|T_D| \times |T_D|}$ are computed as:

$$H, A = \text{PLM}(T_D),$$

where H averages hidden states from the last three PLM layers, and A averages attention weights from all attention heads.

Entity Embedding. For each entity e with mentions $M_e = \{m_i\}_{i=1}^{|M_e|}$, the entity embedding $h_e \in R^d$ is computed using logsumexp pooling over the embeddings of the special tokens at the start of each mention (Jia et al., 2019):

$$h_e = \log \sum_{i=1}^{|M_e|} \exp(H_{m_i}).$$

Localized Context Embedding. ATLOP computes entity-pair localized context embeddings to focus on tokens relevant to both entities in a pair (e_s, e_o). The token importance distribution $q^{(s,o)} \in R^{|T_D|}$ is derived from the cross-token dependencies A:

$$q^{(s,o)} = \frac{a_s \circ a_o}{a_s^\top a_o},$$

where a_s and a_o are the averaged attention weights for entities e_s and e_o, respectively, and $\circ$ denotes the Hadamard product. The localized context embedding $c^{(s,o)} \in R^d$ is then computed as:

$$c^{(s,o)} = H^\top q^{(s,o)}.$$

Relation Classification. For relation classification, ATLOP generates context-aware subject and object representations:

$$\begin{aligned} z_s &= \tanh(W_s[h_{e_s}; c^{(s,o)}] + b_s), \\ z_o &= \tanh(W_o[h_{e_o}; c^{(s,o)}] + b_o). \end{aligned} \tag{4}$$

where $[;]$ denotes concatenation, and $W_s, W_o \in R^{d \times 2d}$ and $b_s, b_o \in R^d$ are trainable parameters. A bilinear classifier computes the relation scores $y^{(s,o)} \in R^{|R|}$:

$$y^{(s,o)} = z_s^\top W_r z_o + b_r,$$

where $W_r \in R^{|R| \times d \times d}$ and $b_r \in R^{|R|}$ are trainable parameters. The probability of relation $r \in R$ is given by $P(r|s,o) = \sigma(y_r^{(s,o)})$, where σ is the sigmoid function.

Loss Function. ATLOP employs Adaptive Thresholding Loss (ATL) to learn a dynamic threshold class TH during training. The loss encourages scores above TH for positive relations R_P and below TH for negative relations R_N:

$$L_{\text{RE}} = -\sum_{s \neq o} \left(\sum_{r \in R_P} \frac{\exp(y_r^{(s,o)})}{\sum_{r' \in R_P \cup \{\text{TH}\}} \exp(y_{r'}^{(s,o)})} - \frac{\exp(y_{\text{TH}}^{(s,o)})}{\sum_{r' \in R_N \cup \{\text{TH}\}} \exp(y_{r'}^{(s,o)})} \right).$$

This approach ensures robust relation classification by adapting to varying document contexts (Chen et al., 2020).

C DREEAM: Guiding Attention with Evidence

Evidence-Guided Supervision. For a given entity pair (e_s, e_o), DREEAM computes an evidence-centered localized context embedding by aggregating token-level attention weights within each sentence. The model is supervised using a human-annotated evidence distribution $v^{(s,o)}$, which guides attention to align with sentence-level evidence. The evidence retrieval (ER) loss minimizes the Kullback-Leibler (KL) divergence between the predicted evidence distribution $p^{(s,o)}$ and the human-annotated distribution:

$$L_{\text{gold}}^{ER} = -D_{KL}(v^{(s,o)} \| p^{(s,o)}). \tag{5}$$

The overall loss combines the relation extraction (RE) loss and the ER loss, weighted by a hyperparameter λ:

$$L_{\text{gold}} = L_{\text{RE}} + \lambda L_{\text{gold}}^{ER}. \tag{6}$$

Teacher-Student Self-Training. DREEAM employs a teacher-student distillation pipeline for self-training on distantly-supervised data (Tan et al., 2022a; Mintz et al., 2009) which contains noisy labels for RE but no information for ER. The teacher model, trained on human-annotated data, predicts evidence distributions for distantly-supervised data, generating silver evidence labels. The student model is then trained to mimic these predictions, using the KL divergence loss:

$$L_{\text{silver}}^{ER} = -D_{KL}(\hat{q}^{(s,o)} \| q^{(s,o)}), \tag{7}$$

where $\hat{q}^{(s,o)}$ is the teacher-predicted evidence distribution, and $q^{(s,o)}$ is the student's evidence prediction.

There are two notable differences between L_{silver}^{ER} and L_{gold}^{ER} .L_{gold}^{ER} employs sentence-level supervision, whereas L_{silver}^{ER} adopts token-level supervision to leverage the fine-grained evidence distribution predicted by the teacher model on distantly supervised data. On the other hand, due to the noisy nature of relation labels in distantly supervised data, L_{silver}^{ER} is computed over all entity pairs, while L_{gold}^{ER} is applied only to entity pairs with valid relations. The final loss follows the same weighting strategy:

$$L_{\text{silver}} = L_{\text{RE}} + \lambda L_{\text{silver}}^{ER}. \tag{8}$$

After self-training, the student model is fine-tuned on human-annotated data to refine its knowledge of both relation extraction and evidence retrieval.

Blending Layer. To further improve relation classification, the model refines relation scores using evidence-based pseudo-documents. A blending layer with a single parameter τ is used to aggregate predictions from the full document and pseudo-documents. A relation triple (e_s, r, e_o) is selected as the final prediction if the summation of its scores from the full document and pseudo-documents exceeds τ. The threshold τ is optimized on the development set to minimize the binary cross-entropy loss of relation extraction.