Document-Level Zero-Shot Relation Extraction with Entity Side Information

Mohan Raj Chanthran[1], **Lay-Ki Soon**[1*], **Ong Huey Fang**[1], **and Bhawani Selvaretnam**[2]
[1]School of Information Technology, Monash University Malaysia
{mohanraj.chanthran, soon.layki, ong.hueyfang}@monash.edu
[2]Valiantlytix
bhawani@valiantlytix.com

Abstract

Document-Level Zero-Shot Relation Extraction (DocZSRE) aims to predict unseen relation labels in text documents without prior training on specific relations. Existing approaches rely on Large Language Models (LLMs) to generate synthetic data for unseen labels, which poses challenges for low-resource languages like Malaysian English. These challenges include the incorporation of local linguistic nuances and the risk of factual inaccuracies in LLM-generated data. This paper introduces Document-Level Zero-Shot Relation Extraction with Entity Side Information (DocZSRE-SI) to address limitations in the existing DocZSRE approach. The DocZSRE-SI framework leverages Entity Side Information, such as Entity Mention Descriptions and Entity Mention Hypernyms, to perform ZSRE without depending on LLM-generated synthetic data. The proposed low-complexity model achieves an average improvement of 11.6% in the macro F1-Score compared to baseline models and existing benchmarks. By utilizing Entity Side Information, DocZSRE-SI offers a robust and efficient alternative to error-prone, LLM-based methods, demonstrating significant advancements in handling low-resource languages and linguistic diversity in relation extraction tasks. This research provides a scalable and reliable solution for ZSRE, particularly in contexts like Malaysian English news articles, where traditional LLM-based approaches fall short.

1 Introduction

Relation Extraction (RE) is a crucial NLP task that identifies the relation between entities in text. Document-Level Relation Extraction (DocRE) takes this further by capturing relations across sentences. Most RE models rely on supervised learning, which requires large amounts of labelled data and is limited to predicting predefined relations. Labelling data is expensive and time-consuming, especially for news articles where new relations constantly emerge. Open Relation Extraction (ORE) tries to address this by identifying relation phrases without predefined labels, but it often produces redundant or overly specific results, making standardization and interpretation difficult. Zero-Shot Relation Extraction (ZSRE) has emerged as a solution to overcome the limitations of supervised and ORE approaches. While significant progress has been made in sentence-level ZSRE, document-level ZSRE remains largely unexplored. Currently, only one notable approach exists for document-level ZSRE (Sun et al., 2024), which generates synthetic data for unseen relations and fine-tunes a language model on seen relations. However, this method is complex, resource-intensive, and relies heavily on LLMs like ChatGPT, which struggle with low-resource languages such as Malaysian English (Chanthran et al., 2023). These limitations highlight the need for more robust and scalable solutions.

To address these gaps, we propose Document-Level Zero-Shot Relation Extraction with Entity Side Information (DocZSRE-SI). This framework tackles the challenges of document-level ZSRE by focusing on Entity Side Information, including Entity Mention Descriptions, Entity Mention Hypernyms, and Entity Types. Instead of processing entire documents, DocZSRE-SI concentrates on the Entity Mention Descriptions relevant to the entity pairs being evaluated. This approach provides a concise and meaningful representation of entities, improving efficiency and accuracy by focusing on essential context and reducing noise from long documents. Our evaluation shows that DocZSRE-SI performs well not only in Malaysian English but also in Standard English. The key contributions of this paper are as follows:

1. Introduction of DocZSRE-SI: We propose Document-Level Zero-Shot Relation Extraction framework that leverages Entity Side In-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4670–4680

formation to enhance the prediction of unseen relations within a document. The code of this framework is published in https://github.com/mohanraj-nlp/DocZSRE-SI.

2. Efficient Context Utilization: Instead of processing entire documents, DocZSRE-SI focuses on relevant Entity Mention Descriptions, reducing noise and improving efficiency while maintaining high accuracy in predicting unseen relations.

3. Incorporation of Entity Side Information: Our approach integrates Entity Mention Hypernym, and Entity Type to provide a richer semantic representation, improving the model's ability to infer relations in a zero-shot setting.

This paper is structured as follows: Section 2 reviews existing approaches for sentence-level and document-level ZSRE. Section 3 presents the DocZSRE-SI framework, which consists of two modules: Building Entity Side Information and Zero-Shot Relation Extraction. In Sections 4 and 5, we describe the experimental setup and discuss the results. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Related Work

ZSRE extracts relations between entities without task-specific labelled data, using pre-trained language models (PLMs), knowledge graphs, or prompt-based methods to generalize to unseen relations. While most ZSRE approaches focus on Sentence-Level (intra-sentential) Relation Extraction (Section 2.1), only one notable work addresses Document-Level (inter-sentential) Relation Extraction (Section 2.2). This section reviews key ZSRE contributions and is grouped by techniques.

2.1 Sentence-Level ZSRE

Sentence-Level Zero-Shot Relation Extraction (ZSRE) has been explored through various methodologies, including reformulating it as Reading Comprehension and Textual Entailment tasks, leveraging Prompt-based Learning with external knowledge, and advancing Representation Learning with matching techniques. Early works framed ZSRE as reading comprehension (Levy et al., 2017) or textual entailment tasks (Obamuyide and Vlachos, 2018), with later improvements like entailment templates (Rahimi and Surdeanu, 2023). Prompt-based methods, such as RelationPrompt (Chia et al., 2022) and ZS-SKA (Gong and Eldardiry, 2021), combined prompts with external knowledge, achieving strong results on datasets like FewRel (Han et al., 2018), and Wiki-ZSL (Chen and Li, 2021) but struggling with generalization to unseen relations. Representation learning approaches, including ZSLRC (Gong and Eldardiry, 2020), ZS-BERT (Chen and Li, 2021), and RE-Matching (Zhao et al., 2023), focused on fine-grained matching and improved zero-shot classification. Additionally, Weak Supervision and Template Infilling have emerged as innovative strategies to reduce reliance on annotated data.

2.2 Document-Level ZSRE

The literature review highlights limited work on Document-Level Zero-Shot Relation Extraction (ZSRE). One notable contribution is by (Sun et al., 2024), which proposes generating synthetic data using ChatGPT to handle unseen relation labels. The approach introduces a Chain-of-Retrieval (CoR) prompt to guide the generation of sentences corresponding to relation triplets and incorporate a Consistency-Guided Knowledge Denoising Strategy to enhance the quality of synthetic data. Experiments show the approach achieves 41.3 ± 8.9 and 41.5 ± 8.7 on Re-DocRED and DocRED test sets, outperforming baselines and demonstrating its effectiveness in generating high-quality relational data without extensive human annotation. Despite this advancement, document-level ZSRE remains underexplored, calling for further research.

3 Methodology

Figure 1 provides an overview of DocZSRE-SI, which consists of two key components. First, the Building Entity Side Information Module processes input document to extract additional details of Entity Mention like Entity Types, Entity Mention Descriptions, and Entity Mention Hypernyms (details Section 3.1). Second, the Zero-Shot Relation Extraction Module analyzes each entity pair to identify the best unseen relation label. This is done by leveraging Entity Side Information and calculating a Dynamic Weighted Score for each label. The label with the highest score is selected as the correct relation (details in Section 3.2).

3.1 Building Entity Side Information

Building Entity Side Information module is a key element of the DocZSRE-SI framework. This component is designed to enrich information about en-

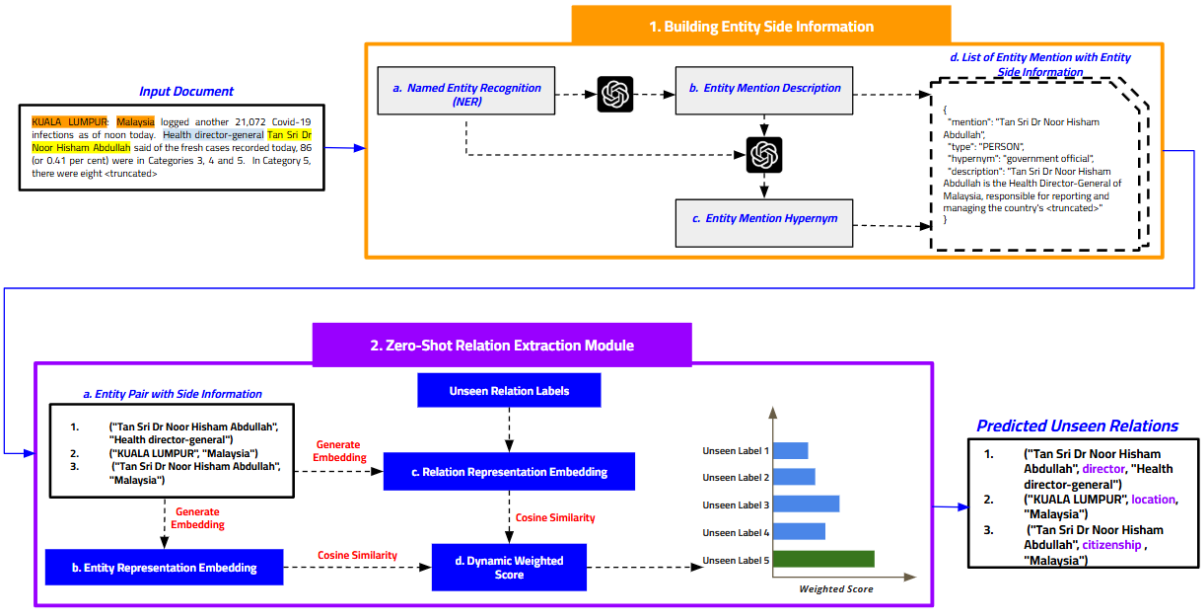

Figure 1: High-Level architecture of DocZSRE-SI Framework

tities based on input documents. The collected side information includes Entity Mention Type, Entity Mention Description (Section 3.1.1), and Entity Mention Hypernym (Section 3.1.2).

3.1.1 Entity Mention Description Generator

Entity Mention Description extends traditional NER by providing rich, contextual details about an entity. Instead of simply labelling it as a *PERSON*, *ORGANIZATION*, or *LOCATION*, it generates a detailed textual description that captures the surrounding text, semantic context, and implied attributes while also pulling information from multiple sentences or paragraphs for a more comprehensive understanding. This approach is used instead of processing the full document because it focuses on the most relevant context for the entity pair, reducing noise and improving efficiency. These descriptions are generated using gpt-4o-mini (OpenAI, 2023), chosen after testing various LLMs for quality. An example of a generated description can be found in Appendix A.

3.1.2 Entity Mention Hypernym Generator

Hypernyms represent the broader category an entity belongs to, providing a higher-level understanding of its classification. Hypernyms can help differentiate entities with the same entity type but different roles. For example, consider these two entity mentions, *Maybank Sdn Bhd* and *Khairussaleh Ramli*. For context, *Khairussaleh Ramli* is the CEO of *Maybank Sdn Bhd*. The entity type of *Maybank Sdn Bhd* is *ORGANIZATION*, and *Khairussaleh Ramli* is *PERSON*, but these labels alone do not provide sufficient context about the entities. By incorporating hypernyms, it becomes more informative that *Maybank Sdn Bhd* is a *banking institution*, and *Khairussaleh Ramli* is a *business executive*. Previous work (Gong and Eldardiry, 2020) showed that hypernyms improve sentence-level ZSRE. These hypernyms are generated by combining entity mention, entity mention type, and entity mention description using gpt-4o-mini.

3.2 Zero-Shot Relation Extraction Module

The Zero-Shot Relation Extraction Module is responsible for identifying unseen relation labels. The module includes components like Entity Side Information Embedding (Section 3.2.1), Relation Representation Embedding (Section 3.2.2), and Dynamic Weighted Score (Section 3.2.4).

3.2.1 Entity Side Information Embedding

The Entity Side Information Embedding component combines the different entity-side features and produces embeddings that are then used in subsequent stages of the framework for ZSRE. These embeddings are derived from different aspects of the entity side information:

1. **Combined Description Embedding**: Merging the descriptions of head and tail entities provides richer context, helping the model understand their roles. However, descriptions

alone may not fully capture the relation between entities.

2. **Entity Hypernym Embeddings**: Hypernyms offer higher-level categories for entities, improving generalization to unseen relations by helping the model recognize broader patterns.

3. **Entity Type Embeddings**: Entity types help differentiate entities by providing basic category information. Combining entity types with hypernyms balances general and specific details, improving relation prediction.

4. **Role-Based Embeddings**: Roles of entity clarify how entities function within a relation by distinguishing between subjects and objects. This prevents the model from misinterpreting entity relationships. Prompt templates for role-based embeddings are defined in Appendix B

5. **Context Embedding**: This component generates an embedding to capture the contextual relation between two entities. It is designed to calculate the cosine similarity with the relation label embedding, aiding in identifying the most likely relation. The prompt used for context embedding is defined in Appendix C. The prompt conveys the idea of a connection between two things, where the head_hypernym and tail_hypernym represent the entities involved. This prompt was included as additional information based on our observation that entity types are often used as arguments when predicting relations.

In this work, we use a pre-trained BERT model (Devlin et al., 2019), specifically `bert-base-uncased` [1], to generate the embeddings. Entity descriptions, types, and hypernyms provide useful context but may not fully capture entity relationships. Descriptions can be vague, hypernyms may overlook unique interactions, and entity types alone lack depth. Combining these with role-based embeddings enhances the model's understanding, thereby improving accuracy in predicting unseen relations.

3.2.2 Relation Label Embedding

Relation Label Embedding converts relation labels into dense vectors, representing unseen relations. We encode each relation label using `bert-base-uncased`[2], following the same encoding strategy used for entity representations. Cosine similarity with Entity Side Information Embeddings helps assess how well an entity pair aligns with a relation. Calculating cosine similarity with Entity Side Information Embeddings helps measure how well the current entity pair aligns with the specific relation. Section 3.2.3 explains this calculation in detail.

3.2.3 Calculating Cosine Similarity

Cosine similarity is used to measure the similarity between the embeddings discussed in Sections 3.2.1 and 3.2.2, allowing the framework to assess how closely the Entity Side Information Embedding matches the target Relation Label Embedding. Cosine similarity measures the angle between two vectors, with values ranging from -1 (completely dissimilar) to +1 (completely similar). The framework computes multiple similarity measures:

1. **Description Similarity Score**: Measures alignment between Relation Label Embedding and Combined Description Embedding.

2. **Entity Hypernym Similarities Score**: Separate similarity scores are calculated between the Relation Label Embedding and the Entity Hypernym Embeddings of the head and tail entities.

3. **Entity Type Similarities Score**: Separate similarity scores are calculated between the Relation Label Embedding and the Entity Type Embeddings of the head and tail entities.

4. **Role-Based Similarities Score**: Equation 1 shows how the role-based similarity score are calculated. Separate scores are calculated between role-based embeddings and relation label context. Role-based embeddings clarify entity roles (subject/object) and resolve ambiguities from types or hypernyms. With two embeddings (head and tail entities), their average is calculated to prevent one entity's information (Type and Hypernym) from heavily influencing the result.

5. **Context Similarity Score**: Evaluates how well the relation between two entity types

[1] https://huggingface.co/google-bert/bert-base-uncased

[2] https://huggingface.co/google-bert/bert-base-uncased

$$\text{Role-Based Score} = \frac{(\text{Head Role-Based Embedding} + \text{Tail Role-Based Embedding})}{2} \quad (1)$$

matches the target label. For instance, a relation between *{person}* and *{educational institution}* might align with labels like *educated_at, place_of_birth*, adding context for better predictions.

6. **Consistency-Based Confidence Weightage**: The confidence score combines the mean and consistency of six similarity measures. Consistency is measured by the standard deviation (lower = more agreement), while mean similarity reflects strength. Averaging these ensures higher confidence when scores are both strong (High Mean) and stable (Low Deviation), making predictions more reliable.

Different similarity measures help the model better understand the relation between entities, especially with unseen relation labels. However, simply summing these scores is not enough for accurate predictions. To improve accuracy, more weight is given to the most important scores, ensuring a stronger influence on the final result (see Section 3.2.4).

3.2.4 Dynamic Weighted Score

Dynamic Weighted Score enhances relation prediction by assigning different importance levels to similarity scores. It prioritizes key features, like entity descriptions, over others (e.g., entity types or hypernyms), ensuring better entity-relation alignment. Weightage will be assigned to each similarity score based on its relevance.

$$\begin{aligned}\text{Dynamic Weighted Score} = \Big(& \\ 0.4 \times \text{Description Similarity (a, r)} & \\ + 0.1 \times \text{Entity Hypernym Similarity (b, r)} & \\ + 0.1 \times \text{Entity Hypernym Similarity (c, r)} & \\ + 0.1 \times \text{Entity Type Similarity (d, r)} & \\ + 0.1 \times \text{Entity Type Similarity (e, r)} & \\ + 0.1 \times \text{Role-Based Similarities Score (f,g,r)} & \\ + 0.1 \times \text{Context Similarity Score} \Big) & \\ \times \text{Consistency-Based Confidence Weightage} & \end{aligned} \quad (2)$$

where:

a = Combined Description Embedding
b = Head Entity Hypernym Embedding
c = Tail Entity Hypernym Embedding
d = Head Entity Type Embedding
e = Tail Entity Type Embedding
f = Head Role-Based Embedding
g = Tail Role-Based Embedding
h = Tail Entity Type Embedding
r = Relation Label Embedding

In Equation 3.2.4, description similarity is weighted highest (0.4), reflecting its importance in capturing entity interactions. Descriptions provide rich contextual details, making them more impactful than other features. We tested weights (0.2, 0.4, 0.6) and chose 0.4 for optimal balance, where lower weights reduced its influence while higher weights overshadowed other features. This ensures description similarity remains significant without diminishing other contributions. The unseen relation label with the highest score is selected, improving prediction accuracy by prioritising contextually relevant features.

4 Experiments

4.1 Evaluation Metrics

ZSRE evaluation follows prior works (Gong and Eldardiry, 2021; Chen and Li, 2021; Chia et al., 2022; Wang et al., 2022; Zhao et al., 2023; Kim et al., 2023; Sun et al., 2024), where unseen relations are randomly chosen from the dataset. Unseen relation sets of sizes $n \in \{5, 10, 15\}$ are used, with three random samples for each size to ensure results aren't biased by specific relation choices. This random sampling tests the model's generalization across different scenarios. Macro F1-Score is the primary metric, calculated for each run and averaged, with variance reported to measure consistency. Low variance indicates stable and reliable performance, while high variance suggests sensitivity to specific data samples. This setup ensures a fair and robust evaluation.

4.2 Dataset and Benchmarking

This research uses DocRE datasets, including MEN-Dataset (Chanthran et al., 2024), DocRED (Yao et al., 2019), and RE-DocRED (Tan et al., 2022). We split our experiments into two parts. The first part consists of an ablation study conducted on the MEN dataset, RE-DocRED, and a subset of the DocRED dataset, where only 20% of the documents (21,577 documents) are used due to computational constraints. The second part evaluates the model using the full development and test sets of DocRED and RE-DocRED, respectively. Unseen relation labels will be randomly selected for fair evaluation. A baseline will be established using only Entity Mention Descriptions, excluding features such as Entity Type, Hypernym, or Dynamic Weighted Score. The proposed framework will be compared with existing document-level ZSRE methods, particularly (Sun et al., 2024).

4.3 Experiments Planned

To evaluate DocZSRE-SI, we conduct two key experiments. First, an ablation study examines the impact of several key features, including Entity Mention Descriptions, Entity Mention Types, Entity Mention Hypernyms, and a Weighted Dynamic Scoring mechanism. This helps us to better understand how each feature contributes to the overall performance of our approach and identify the optimal combination for improving prediction accuracy. The results of this experiment are presented in Table 1. As this is a ZSRE, we conduct experiments on the full MEN dataset, the RE-DocRED dataset, and 20% of the DocRED dataset. Second, we compare our approach with GenRDK, a method proposed by (Sun et al., 2024) for predicting unseen relations. Performance is evaluated on the dev and test sets of DocRED and RE-DocRED. The unseen relation labels are randomly selected to ensure fairness. This comparison offers insights into the effectiveness of our framework in comparison to existing Document-Level ZSRE methods.

5 Results and Discussion

5.1 Ablation Study

Various Entity Side Information was introduced to enhance the proposed methodology. An ablation study was conducted on three different datasets to evaluate its effectiveness. The study compared five approaches for predicting unseen relations, using the macro F1-Score as the primary metric. Table

		RE-DocRED	DocRED	MEN-Dataset
5	Only Entity Mention Description (Baseline)	28.14 ± 5.43	36.34 ± 5.43	27.65 ± 24.46
	Entity Mention Description + Entity Mention Hypernym	42.34 ± 2.39	42.23 ± 2.39	33.96 ± 4.88
	Entity Mention Description + Entity Mention Type	35.12 ± 5.65	31.41 ± 5.65	28.33 ± 15.76
	Entity Mention Description + Entity Mention Hypernym + Entity Mention Type	38.67 ± 8.1	36.44 ± 8.15	39.79 ± 10.66
	Entity Mention Description + Entity Mention Hypernym + Entity Mention Type + Dynamic Weighted Score (Proposed Approach)	50.05 ± 8.37	48.83 ± 7.57	40.25 ± 7.53
10	Only Entity Mention Description (Baseline)	20.47 ± 5.81	22.36 ± 5.81	19.89 ± 18.19
	Entity Mention Description + Entity Mention Hypernym	39.89 ± 12.79	37.7 ± 12.79	30.75 ± 10.95
	Entity Mention Description + Entity Mention Type	25.82 ± 6.35	22.67 ± 6.35	23.49 ± 6.52
	Entity Mention Description + Entity Mention Hypernym + Entity Mention Type	35.83 ± 14.2	33.59 ± 14.1	26.13 ± 9.32
	Entity Mention Description + Entity Mention Hypernym + Entity Mention Type + Dynamic Weighted Score (Proposed Approach)	44.98 ± 6.78	43.43 ± 8.64	36.02 ± 8.22
15	Only Entity Mention Description (Baseline)	8.56 ± 5.05	8.4 ± 5.05	16.39 ± 4.84
	Entity Mention Description + Entity Mention Hypernym	31.66 ± 9.77	31.02 ± 9.77	22.74 ± 4
	Entity Mention Description + Entity Mention Type	16.78 ± 3.88	15.41 ± 3.88	22.64 ± 6.22
	Entity Mention Description + Entity Mention Hypernym + Entity Mention Type	31.67 ± 4.25	29.76 ± 4.11	23.11 ± 2.07
	Entity Mention Description + Entity Mention Hypernym + Entity Mention Type + Dynamic Weighted Score (Proposed Approach)	33.91 ± 4.57	32.55 ± 4.57	24.28 ± 5.1

Table 1: The complete result of the Ablation Study was conducted for the Zero-Shot Relation Extraction Module for DocRED, RE-DocRED, and MEN-Dataset. Reported in this result are F1-Score with the Variance.

1 shows the performance of each approach based on Macro F1-Score together with Variance. Further analysis has been conducted to compare the performance of various approaches and the impact of Entity Side Information and Dynamic Weighted Scores.

Performance Comparison Across Different Approaches

The proposed approach, combining Entity Mention Hypernym, Entity Mention Type, and Dynamic Weighted Score, achieved the highest macro F1-Score across three datasets (RE-DocRED, DocRED, MEN-Dataset). The proposed approach outperformed the Baseline (using only Entity Mention Descriptions) by 164.56% (RE-DocRED), 83.21% (DocRED), and 35.95% (MEN-Dataset). It correctly predicted 90% of unseen labels, with DocRED showing the highest accuracy improvements (15-30%) and MEN-Dataset the lowest, likely due to its smaller size and fewer instances. Low variance (e.g., 6.57 for RE-DocRED) indicated robustness across unseen relation groups, demonstrating consistent and accurate results.

Impact of Different Entity Side Information

Each type of Entity Side Information contributes differently to performance. The Baseline uses only Entity Mention Descriptions, replacing full documents with combined descriptions. The ablation study evaluates the impact of adding Entity Mention Type and Entity Mention Hypernym, individually and combined.

Results in Table 1 show that combining Entity Mention Descriptions with Hypernyms consistently outperforms combining descriptions with Types, with performance gaps of 45.63% (RE-DocRED), 52.43% (DocRED), and 18.13% (MEN-Dataset). This highlights hypernyms' stronger contribution, providing broader semantic understanding and aiding generalization across unseen relations. In contrast, Entity Mention Types, while useful, are limited by their high-level categorical nature. Interestingly, combining all three (descriptions, types, and hypernyms) did not improve performance, as too many similarity scores introduced noise. This led to the development of a Dynamic Weighting Mechanism (Equation 3.2.4), which prioritizes significant information sources and boosts prediction confidence by assigning appropriate weights to similarity measures.

Impact of Dynamic Weighted Score

After observing a decrease in performance, we incorporated Weighted Score and Confidence into Equation 3.2.4. This significantly improved macro F1-Score. Too many similarity scores overshadowed relevant features, so higher weight was assigned to Entity Mention Descriptions due to their importance, while other side information received equal weight.

However, assigning weights alone didn't fully address discrepancies from multiple side information sources. To tackle this, a Confidence Score was introduced, calculated using the mean and standard deviation of similarity scores. A high mean indicates strong alignment, while a low standard deviation ensures agreement among scores, enhancing reliability. For each entity pair, the highest similarity score determines the correct unseen relation label. Inconsistent measures (high standard deviation) highlight noise, and the confidence score mitigates this, improving prediction reliability. As shown in Table 1, the Best Approach (Entity Mention Description + Hypernym + Type + Dynamic Weighted Score) outperformed the approach without dynamic weighting by 20.62% (RE-DocRED), 24% (DocRED), and 14.84% (MEN-Dataset). This demonstrates the effectiveness of incorporating contextual information and dynamic weighting.

5.2 Comparing with Benchmark Approach

In Section 5.1, we compared different approaches to highlight the impact of Entity Side Information on predicting unseen relation labels. To assess relative performance, we compared our Best Approach (Entity Mention Description + Hypernym + Type + Dynamic Weighted Score) with GenRDK, the only existing Document-Level ZSRE method (Sun et al., 2024). Table 2 shows the comparison based on macro F1-Score.

Our Best Approach outperformed GenRDK, with an average improvement of 40.04% for $n \in \{10\}$ in RE-DocRED and DocRED. However, for $n \in \{5\}$, the improvement was smaller (12.83%), and no comparison was possible for $n \in \{15\}$ due to missing GenRDK results. This suggests our approach excels with larger unseen label groups but shows diminishing gains with fewer labels. For $n \in \{5\}$, our approach had a lower variance than GenRDK, indicating more stable predictions. However, for $n \in \{10\}$, while achieving a higher macro F1-Score, our approach exhibited higher variance, suggesting sensitivity to specific relation labels.

n	5						10						15			
Approach	GenRDK		Best Approach**		% Improvement		GenRDK		Best Approach**		% Improvement		GenRDK		Best Approach**	
Dataset	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Re-DocRED	39.9 ± 10.9	41.3 ± 8.9	**50.95 ± 10.2**	**47.81 ± 9.78**	+27.69	+15.76	30.6 ± 3.6	30.1 ± 4.2	**47.7 ± 12.71**	**43.06 ± 15.5**	+55.8	+43.05	-	-	**30.87 ± 4.4**	**33.88 ± 5.55**
DocRED	42.5 ± 10.6	41.5 ± 8.7	**45.12 ± 7.31**	-	+6.34	-	33.7 ± 4.0	31.4 ± 4.6	**40.51 ± 12.65**	-	+20.20	-	-	-	**31.68 ± 6.4**	-

Table 2: Comparing the macro F1-Score of GenRDK and our **Best Approach (Entity Mention Description + Entity Mention Hypernym + Entity Mention Type + Dynamic Weighted Score). The DocRED dataset provides a blind test set, which prevented us from directly evaluating the performance of the test set. Result for GenRDK taken from (Sun et al., 2024)

Detailed analysis revealed that one relation label group (r=3) had significantly lower F1-Scores and higher variance, likely due to fewer relation instances, semantic ambiguity, or errors in side information (e.g., entity types or hypernyms). This highlights the need for further investigation into label complexity and data quality.

5.3 Exploring Inter and Intra-Sentential Relation Extraction Capabilities

The proposed Document-Level ZSRE uniquely uses Entity Mention Descriptions instead of entire documents to predict relations. These descriptions capture the context of entity mentions from the original news article, with the head and tail entity descriptions concatenated as input. To ensure this approach doesn't impact the performance, results were analyzed by comparing MEN-Dataset and RE-DocRED using relation label lists $n \in \{5, 10\}$. Since RE-DocRED includes the full dataset (unlike DocRED, which only has 20%), it was used in the experiments.

In Appendix D, we present the results of an experiment conducted to evaluate the effectiveness of DocZSRE-SI in performing both inter-sentential and intra-sentential ZSRE. For both datasets (MEN-Dataset and RE-DocRED), the highest accuracy is observed when the sentence gap is 0, indicating that intra-sentence relations are easier to identify. Accuracy drops as the sentence gap increases in both datasets, suggesting that inter-sentence relations are harder to capture, for the MEN-Dataset at $n \in \{5\}$, the accuracy declines from 56.69% (gap = 1) to 16.93% (gap $\geq$ 5). Meanwhile for RE-DocRED at $n \in \{5\}$, the accuracy drops from 54.02% (gap = 1) to 46.69% (gap $\geq$ 5). Incorrect predictions increase significantly for larger sentence gaps.

The results clearly show that sentence gaps negatively affect relation extraction performance. This is especially pronounced for inter-sentence relations (gap $\geq$ 1), where accuracy consistently declines as the gap widens. For $n \in \{10\}$, the percentage of correct is lower than incorrect, as discussed in the previous section. This analysis demonstrates that while intra-sentence relations are well-handled by the proposed solution, inter-sentence relations remain a significant challenge when the gaps increase. Addressing this gap will require advanced modelling techniques, improved datasets, and leveraging external knowledge sources.

6 Conclusion

This paper introduces the Document-Level Zero-Shot Relation Extraction with Entity Side Information (DocZSRE-SI) framework, designed to predict unseen relation labels, especially in dynamic scenarios with emerging relations. A core component is Entity Side Information, which generates features like Entity Mention Descriptions, Types, and Hypernyms to provide contextual details for accurate predictions. Using Entity Side Information instead of the full document allows the model to focus on meaningful features of entities, reducing noise and improving generalization to unseen relations. Evaluations on standard datasets, such as DocRED and RE-DocRED, and language-specific datasets, like MEN-Dataset, demonstrate the framework's adaptability. Although the Best Approach shows promising results, the high variability in performance compared to the benchmark is worth further investigation in the future. Despite this, the work presented in this chapter provides a starting point for improving Document-Level ZSRE methods for real-world use. For future work, we plan to enhance the Zero-Shot Relation Extraction Mod-

ule to reduce variance and expand the application of Entity Mention Side Information to Supervised DocRE.

7 Limitations

While our proposed framework, DocZSRE-SI, outperforms existing Document-Level ZSRE, it shows higher variance in certain scenarios, such as when $n \in \{10\}$. Our investigation revealed that the approach is sensitive to semantically similar relation labels. Additionally, relying on entity types introduces ambiguity when using generic types like `MISC`, highlighting the need for more specific and accurate entity type assignments. Finally, the framework is simpler than some LLM-based approaches, but it involves multiple Similarity Score calculations and Dynamic Weighted Scores, which are computationally intensive for large-scale applications. We also acknowledge the concern about relying on manually chosen parameters or coefficients that may require adjustment for different applications or use cases. However, in our approach, the coefficients used for combining similarity scores are fixed and are selected after a few iterations of preliminary experiments on a validation set, rather than being tuned for each dataset or domain.

8 Acknowledgement

We would like to acknowledge the responsible use of Generative AI tools, which assisted in error checking and improving the clarity of my writing in compliance with academic integrity guidelines. Large Language Models (LLMs) like gpt-4o-mini were employed to generate Entity Mention Descriptions and Hypernyms. Part of this project was funded by the Malaysian Fundamental Research Grant Scheme (FRGS) FRGS/1/2022/ICT02/MUSM/02/2.

References

Mohanraj Chanthran, Lay-Ki Soon, Ong Huey Fang, and Bhawani Selvaretnam. 2023. How well ChatGPT understand Malaysian English? an evaluation on named entity recognition and relation extraction. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 372–397, Singapore. Association for Computational Linguistics.

MohanRaj Chanthran, Lay-Ki Soon, Huey Fang Ong, and Bhawani Selvaretnam. 2024. Malaysian English news decoded: A linguistic resource for named entity and relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10999–11022, Torino, Italia. ELRA and ICCL.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland. Association for Computational Linguistics.

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.

Jiaying Gong and Hoda Eldardiry. 2020. Zero-shot relation classification from side information. *Preprint*, arXiv:2011.07126.

Jiaying Gong and Hoda Eldardiry. 2021. Prompt-based zero-shot relation extraction with semantic knowledge augmentation. In *International Conference on Language Resources and Evaluation*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Bosung Kim, Hayate Iso, Nikita Bhutani, Estevam Hruschka, Ndapa Nakashole, and Tom Mitchell. 2023. Zero-shot triplet extraction by template infilling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 272–284, Nusa Dua, Bali. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4. https://arxiv.org/abs/2303.08774.

Mahdi Rahimi and Mihai Surdeanu. 2023. Improving zero-shot relation classification via automatically-acquired entailment templates. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 187–195, Toronto, Canada. Association for Computational Linguistics.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4407–4416, New York, NY, USA. Association for Computing Machinery.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. 2022. RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, United States. Association for Computational Linguistics.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Jun Zhao, WenYu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng, and Mingming Sun. 2023. RE-matching: A fine-grained semantic matching method for zero-shot relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6680–6691, Toronto, Canada. Association for Computational Linguistics.

A Example of Generated Entity Mention Description

Consider the following news article snippet as an example:

KUALA LUMPUR: Malaysia logged another 21,072 Covid-19 infections as of noon today. Health director-general Tan Sri Dr Noor Hisham Abdullah said of the fresh cases recorded today, 86 (or 0.41 per cent) were in Categories 3, 4 and 5. In Category 5, there were eight cases (0.04 per cent); Category 4 with 20 cases (0.09 per cent), and Category 3 with 58 (0.28per cent) ....

To generate a description for Tan Sri Dr. Noor Hisham Abdullah (PERSON), the content will be generated based on the entity mention. The generated description is:

Tan Sri Dr Noor Hisham Abdullah is the Health Director-General of Malaysia, responsible for reporting and managing the country's Covid-19 response during the pandemic.

This description provides contextual information about the entity, helping to enhance the model's understanding of its role and significance within the document.

B Prompt Template for Role-Based Embeddings

- Head: `{head_type} acting as a subject, described as {head_hypernym}`
- Tail: `{tail_type} acting as a subject, described as {tail_hypernym}`

C Prompt Template for Context Embeddings

- `Relation between {head_hypernym} and {tail_hypernym}.`

D Result of Experiment to Evaluate the Inter and Intra-Sentential Capabilities of DocZSRE-SI

Table 3 shows the result of our analysis to understand the effectiveness of our approach when handling both inter and intra-sentential relation extraction.

	MEN-Dataset						RE-DoCRED					
	5			10			5			10		
Sentence Gap (H-T)	Total Instance	Correct (%)	Incorrect (%)	Total Instance	Correct (%)	Incorrect (%)	Total Instance	Correct (%)	Incorrect (%)	Total Instance	Correct (%)	Incorrect (%)
0*	465	50.65	49.35	1463	57.5	42.5	3277	59.05	40.95	5135	32.99	67.01
1	356	56.69	43.31	795	40.5	59.5	1251	54.02	45.98	3043	30.89	69.11
2	206	45.22	54.78	495	26.26	73.74	548	50.55	49.45	711	39.24	60.76
3	54	41.11	58.89	132	37.58	62.42	405	49.14	50.86	459	37.69	62.31
4	32	30.38	69.62	105	30.14	69.86	373	48.26	51.74	397	31.49	68.51
≥ 5	189	16.93	83.07	485	11.75	88.25	829	46.69	53.31	976	30.74	69.26

Table 3: The overall count of sentence gaps between head and tail entities and the percentage of correct and incorrect predictions for each gap category (0, 1, 2, 3, 4, and ≥ 5). *0 refers to condition where Head Entity and Tail Entity are in same sentence (Intra-Sentential RE).