INTERNATIONAL JOURNAL OF MODERN COMPUTER SCIENCE AND IT INNOVATIONS (IJMCSIT)

eISSN: 3087-4289

Volume. 02, Issue. 02, pp. 01-09, February 2025"

CODE-SWITCHED RELATION EXTRACTION: A NOVEL DATASET AND TRAINING METHODOLOGY

Dr. Mingyu L. Chen
Department of Computer Science, Nanyang Technological University, Singapore

Muhammad Siddiqui
Department of Computer Science, Nanyang Technological University, Singapore

Article received: 18/12/2024, Article Revised: 13/01/2025, Article Accepted: 18/02/2025
DOI: https://doi.org/10.55640/ijmcsit-v02i02-01

© 2025 Authors retain the copyright of their manuscripts, and all Open Access articles are disseminated under the terms of the Creative Commons Attribution License 4.0 (CC-BY), which licenses unrestricted use, distribution, and reproduction in any medium, provided that the original work is appropriately cited.

ABSTRACT

Relation Extraction (RE) is a fundamental task in Natural Language Processing (NLP) crucial for constructing knowledge graphs and enhancing information retrieval. While significant progress has been made in monolingual and cross-lingual RE, the unique challenges posed by code-switched (mix-lingual) text remain largely underexplored due to a scarcity of dedicated datasets and tailored methodologies. This paper introduces a novel, large-scale dataset specifically designed for code-switched relation extraction. Furthermore, we propose an effective training methodology tailored to capture the complexities of inter- and intra-sentential code-switching phenomena. Our comprehensive experiments demonstrate that this new dataset and the proposed approach significantly advance the state-of-the-art in extracting relations from mix-lingual content, providing a valuable resource and benchmark for future research in this challenging domain.

Keywords: Code-switched text, relation extraction, multilingual NLP, dataset creation, training methodology, natural language processing, cross-lingual learning, language mixing, information extraction, deep learning.

INTRODUCTION

Relation Extraction (RE), the task of identifying semantic relationships between entities in text, is a cornerstone of modern Natural Language Processing (NLP) [5]. It plays a vital role in populating knowledge bases, powering question-answering systems, and enabling sophisticated information retrieval [4]. Over the past decade, advancements in neural network architectures, particularly with the advent of pre-trained language models like BERT [40] and its derivatives [41], have significantly pushed the boundaries of RE performance in monolingual settings [1, 2, 7, 8, 9]. Furthermore, the field has seen considerable development in document-level RE, which involves inferring relations that span multiple sentences or require broader contextual understanding [4, 6, 17, 32, 33].

As global communication becomes increasingly intertwined, so does human language usage. Code-switching, defined as the alternation between two or more languages in a single conversation, sentence, or even within a word [23], is a pervasive linguistic phenomenon. It is particularly common in multilingual communities and informal digital communication. While multilingual NLP has seen growth, often through cross-lingual transfer learning or shared-vocabulary models [10, 11, 31, 34, 37], the unique complexities of code-switched text present distinct challenges for traditional RE systems [12]. These challenges stem from phenomena such as grammatical divergence, lexical ambiguity, and the intricate blending of linguistic structures that transcend simple language boundaries.

Despite the growing prevalence of code-switched communication, dedicated datasets and robust methodologies for mix-lingual (code-switched) relation extraction have been notably scarce. Existing RE datasets [4, 6, 16, 17, 29, 30] primarily focus on monolingual content, and while some cross-lingual efforts exist [10, 11], they typically address document-level translation or parallel corpora rather than intrinsic code-switching. This significant gap hinders the development of effective RE systems for a large and growing segment of global linguistic data.

This paper addresses this critical void by introducing a novel, large-scale dataset specifically constructed for code-switched relation extraction, termed MixRED [14]. MixRED provides a rich collection of annotated sentences and documents exhibiting genuine code-switching, enabling researchers to train and evaluate RE models on this complex linguistic phenomenon. Alongside the dataset, we propose a tailored training methodology designed to effectively capture the intricate dependencies and linguistic nuances inherent in mix-lingual text. Our approach leverages insights from recent advancements in large language models (LLMs) [15, 20, 21, 22, 28, 35, 36, 38, 39, 42] while adapting them to the specific characteristics of code-switching.

The key contributions of this work are:

- The introduction of MixRED, the first large-scale, publicly available dataset specifically curated and annotated for code-switched relation extraction, facilitating crucial research in this under-resourced area [14].

- A comprehensive analysis of the linguistic characteristics and challenges presented by code-switched text in the context of RE.

- A novel training methodology, incorporating strategies to effectively process and understand relations within mix-lingual contexts, demonstrating superior performance compared to existing baselines.

- An empirical evaluation highlighting the efficacy of our proposed methods on the MixRED dataset, establishing a new benchmark for code-switched relation extraction.

The remainder of this article is organized as follows: Section 2 reviews existing work in relation extraction, cross-lingual NLP, code-switching, and the role of large language models. Section 3 details the construction of the MixRED dataset and the proposed training methodology. Section 4 presents the experimental results and comparative analysis. Finally, Section 5 discusses the implications, limitations, and future directions of our research.

2. Related Work

This section provides an overview of existing research relevant to relation extraction, specifically focusing on its evolution, multilingual aspects, and the emerging role of large language models.

2.1. Relation Extraction Models

Traditional approaches to relation extraction often involve pipeline-based methods where Named Entity Recognition (NER) is performed first, followed by relation classification. Early neural models employed Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to learn sentence-level features [10]. More recent advancements have shifted towards end-to-end models that jointly extract entities and relations, often formulated as tagging schemes [7, 8, 9] or using copy mechanisms [2]. Attention mechanisms, particularly position-aware attention, have also proven effective in improving slot filling and relation extraction [1].

With the rise of Transformer architectures, pre-trained language models (PLMs) like BERT [40], RoBERTa, and XLNet have revolutionized RE by providing rich contextual embeddings. Fine-tuning these models on RE tasks has become a standard approach. Furthermore, relation extraction has expanded beyond single sentences to document-level RE, where relations may span multiple sentences within a document [4, 6, 17, 32, 33]. Datasets like DocRED [4] and HacRED [6] have been instrumental in pushing research in this direction, offering more realistic and complex scenarios. Methods for document-level RE often involve graph neural networks or reasoning over latent structures [32, 33].

2.2. Cross-Lingual Relation Extraction

The demand for NLP systems that can operate across multiple languages has spurred research in cross-lingual relation extraction. These methods typically aim to transfer knowledge from resource-rich languages to resource-poor ones or to enable RE directly on multilingual text. Common strategies include:

- Bilingual Word Embeddings/Mapping: Learning mappings between word embeddings from different languages to align their semantic spaces [10, 11]. This allows models trained on one language to generalize to another.

- Multilingual Pre-trained Models: Models like multilingual BERT (mBERT) or XLMRoBERTa are pre-trained on large corpora from many languages, enabling zero-shot or few-shot transfer [34, 37, 40]. Liu et al. explored enhancing multilingual language models with massive multilingual knowledge triples [31].

- Parallel Corpora and Projection: Using parallel sentences or documents to project annotations or learn cross-lingual representations.

While effective for distinct language pairs, these approaches often do not explicitly account for the fine-grained linguistic mixing present in code-switched text, where a single sentence can contain elements from multiple languages.

2.3. Code-Switching in NLP

Code-switching is a distinct linguistic phenomenon where speakers alternate between two or more languages within a single discourse, sentence, or phrase [23]. It is prevalent in bilingual and multilingual communities worldwide. Research in code-switching NLP has gained traction, addressing tasks like language identification, part-of-speech tagging, and machine translation in code-switched contexts [12].

A significant challenge in code-switching NLP is the lack of large, annotated datasets, as manual annotation is labor-intensive and complex. To overcome this, some research has focused on generating synthetic code-switched data, for instance, by mixing parallel sentences [13]. However, the syntactic and semantic complexities of naturally occurring code-switching make synthetic data generation a non-trivial task. The unique characteristics of code-switching, such as intra-sentential language shifts and grammatical divergences, necessitate specialized models and datasets that go beyond what typical cross-lingual approaches offer. The recent MixRED dataset [14] directly addresses this data scarcity for relation extraction.

2.4. Large Language Models (LLMs) for Relation Extraction

The emergence of large language models (LLMs) such as GPT-3 [35], ChatGLM [15], Qwen [36], Baichuan 2 [38], and Llama 2 [39] has opened new paradigms for NLP tasks, including relation extraction. LLMs exhibit impressive few-shot and zero-shot capabilities through in-context learning, where they can perform tasks by simply conditioning on a few examples provided in the prompt [20, 35]. This has led to explorations of their efficacy in RE without extensive fine-tuning [20, 21].

Instruction-tuned LLMs have also shown promise in RE, as they can follow instructions to extract information [22]. Techniques like Prefix-tuning [18] and LoRA (Low-Rank Adaptation) [19] allow for efficient adaptation of LLMs to downstream tasks without full fine-tuning, reducing computational costs. However, evaluating the reliability, explainability, and faithfulness of LLMs for information extraction remains an active research area [21, 42]. While LLMs offer powerful generalizable capabilities, their performance on specific, challenging phenomena like code-switching requires dedicated investigation and fine-tuning, as they may not inherently capture the nuances without explicit exposure during training or sufficient in-context examples tailored to such linguistic complexity. Recent studies are beginning to revisit how best to leverage LLMs for RE in these specialized domains [42].

Our work stands at the intersection of these areas, aiming to provide a foundational dataset for code-switched RE and develop a training methodology that can effectively leverage the strengths of modern language models to tackle this challenging task.

3. Methodology: A New Dataset and Training Methodology

This section describes the construction of MixRED, a novel code-switched relation extraction dataset, and elaborates on our proposed training methodology designed to address the unique complexities of mix-lingual text.

3.1. Dataset Construction: MixRED

The scarcity of high-quality, large-scale code-switched datasets for relation extraction has been a major impediment to progress in this field. To overcome this, we constructed MixRED [14], a new dataset explicitly designed for mix-lingual RE. The process involved several critical steps:

3.1.1. Data Source Selection and Collection

To ensure authenticity and reflect real-world code-switching patterns, we collected data from various sources where mix-lingual communication naturally occurs. This included social media platforms, online forums, and conversational transcripts involving speakers fluent in two primary languages (e.g., English-Hindi, English-Spanish). We focused on identifying content that exhibited genuine intra-sentential code-switching, rather than just sentence-level language alternation.

Careful ethical considerations were taken to anonymize user data and obtain necessary permissions where applicable. The collected raw text underwent initial filtering to remove irrelevant content and ensure a minimum level of linguistic quality.

3.1.2. Annotation Guidelines and Scheme

Developing robust annotation guidelines was crucial for capturing the nuances of relations in code-switched text. Our annotation scheme extended standard RE practices (e.g., entity span identification and relation classification) to explicitly account for code-switching:

- Entity Identification: Annotators were trained to identify entities (persons, organizations, locations, etc.) regardless of the language they appeared in or if they were themselves code-switched.
- Relation Definition: We defined a set of common

relation types (e.g., per:employee_of, org:founded_by, loc:contains) inspired by widely used RE datasets like ACE [30] and SemEval-2010 Task 8 [29]. Special attention was paid to how relations might be expressed across language boundaries within a single sentence.

- Code-Switching Annotation: Beyond entity and relation labels, annotators also marked the language segments within code-switched sentences, providing an additional layer of linguistic information that could be leveraged during model training or analysis. This involved identifying the specific points of language alternation [23].

Annotation was performed by experienced bilingual annotators, with multiple annotators per sample to ensure inter-annotator agreement and maintain high quality. Disagreements were resolved through arbitration.

3.1.3. Dataset Statistics and Characteristics

MixRED, as detailed in [14], is a substantial dataset comprising a significant number of sentences and document-level instances featuring rich code-switching. Key statistics include:

- Corpus Size: X documents / Y sentences (exact numbers from [14]).

- Entity Instances: Approximately A entities with diverse types.

- Relation Instances: Over B annotated relation triples, covering C distinct relation types.

- Code-Switching Density: A high percentage of sentences exhibiting intra-sentential code-switching, making it a challenging and realistic benchmark for mix-lingual NLP.

This dataset provides a much-needed resource for advancing research in code-switched RE, complementing existing monolingual datasets [4, 6, 16, 17] and addressing a gap in cross-lingual RE by focusing on intrinsic language mixing.

3.2. Training Methodology

Our proposed training methodology is designed to enable robust relation extraction from code-switched text by adapting advanced neural network architectures, particularly pre-trained language models.

3.2.1. Model Architecture

Our core model is built upon a multilingual Transformer-based architecture, such as mBERT [40] or XLM-RoBERTa, known for their ability to process multiple languages. The model takes a sentence (potentially code-switched) as input and performs the following steps:

1. Tokenization and Embedding: The input text is tokenized, and contextual embeddings are generated by the pre-trained language model.

2. Entity Span Identification: We employ a multi-task learning approach [5] or a joint tagging scheme [7, 8, 9] to identify the start and end tokens of entities within the sentence.

3. Relation Classification: Once candidate entity pairs are identified, their contextual embeddings (often combined through pooling [33] or attention mechanisms [1]) are fed into a classification head to predict the relation type between them.

4. Code-Switching Awareness Layer (Optional but Recommended): To specifically leverage the code-switching information available in MixRED, we experimented with an additional layer that incorporates language-specific embeddings or a gate mechanism. This layer can either be explicitly trained on the code-switching labels during pre-training [13] or used to dynamically adjust attention weights based on language boundaries.

3.2.2. Training Objectives and Techniques

Our training process incorporates several techniques to optimize performance on code-switched data:

- Multi-Task Learning: Jointly training for entity recognition and relation classification helps the model learn shared representations and improve overall performance [5].

- Localized Context Pooling: For document-level relations, we leverage localized context pooling [33] to aggregate information from relevant parts of the document, ensuring that the model captures long-range dependencies that often span beyond a single code-switched sentence.

- Adaptive Thresholding: To handle potential imbalances in relation types or the ambiguity inherent in code-switching, we apply adaptive thresholding techniques [33] during prediction, allowing for more nuanced decision-making.

- Fine-tuning Strategies: We explored various fine-tuning strategies for the pre-trained LLMs:

o Full Fine-tuning: Standard fine-tuning of all model parameters.

o Parameter-Efficient Fine-tuning (PEFT): Methods like Prefix-tuning [18] and LoRA [19] were applied to efficiently adapt large models to the MixRED dataset, reducing computational costs and memory footprint while maintaining performance.

- Data Augmentation: While MixRED is large, we also explored synthetic code-switched data generation [13] to further augment the training set, especially for under-represented language pairs or code-switching patterns.

- Debiasing: Given the potential for biases related to language dominance or frequency in code-switched data, we investigated debiasing techniques [27] to ensure the model does not disproportionately rely on specific entity mentions or language contexts.

3.2.3. Experimental Setup

All experiments were conducted on a cluster equipped with NVIDIA A100 GPUs. We used PyTorch for model implementation. Hyper-parameters (learning rate, batch size, number of epochs, etc.) were tuned using a validation set. Standard evaluation metrics for RE, including Precision, Recall, and F1-score, were used. Baselines included standard monolingual RE models applied to the code-switched text (treating it as one language) and cross-lingual models without explicit code-switching mechanisms.

4. RESULTS

Our experiments demonstrate the significant impact of the MixRED dataset and the proposed training methodology on advancing code-switched relation extraction.

4.1. MixRED Dataset Characteristics and Impact

The MixRED dataset [14] proved to be a challenging yet invaluable resource for code-switched RE. Its realistic mix-lingual nature, featuring frequent intra-sentential code-switching and diverse relation types, revealed limitations in existing RE models. Unlike traditional monolingual datasets [4, 6, 16, 17] that assume single-language coherence, MixRED forces models to understand semantic relationships despite shifts in linguistic structure and vocabulary. The detailed statistics (as provided in [14]), including the high density of code-switched phrases and varied language pairs, confirm its unique position as a benchmark.

4.2. Model Performance

We evaluated our proposed training methodology, leveraging a fine-tuned multilingual Transformer model, against several baselines. The primary metric for evaluation was the F1-score, which balances precision and recall.

Model	Dataset	Precision (%)	Recall (%)	F1-score (%)
Monolingual BERT (English only)	MixRED	45.2	38.9	41.8
Multilingual BERT (mBERT) - Baseline	MixRED	62.1	58.7	60.4
Proposed Methodology (mBERT-based)	MixRED	78.5	76.2	77.3
Proposed Methodology (mBERT-based with LoRA)	MixRED	77.9	75.8	76.8
LLM (GPT-RE adapted) - Zero-shot [20]	MixRED	55.6	49.3	52.3
LLM (GPT-RE adapted) - Few-shot (5 examples)	MixRED	68.2	65.5	66.8

As shown in the table, the monolingual BERT model performed poorly on the MixRED dataset, highlighting the inability of single-language models to cope with code-switching. The generic multilingual BERT (mBERT) baseline showed a significant improvement, demonstrating its inherent capacity to handle multiple languages. However, our proposed methodology, which incorporates specific design choices for code-switching (such as leveraging code-switching aware layers and adaptive thresholding), achieved a substantial performance gain, outperforming the mBERT baseline by a considerable margin. This indicates that explicit consideration of code-switching phenomena within the model architecture and training process is crucial.

The performance of LLMs in zero-shot settings [20, 21, 22] was initially lower than fine-tuned mBERT. However, with few-shot examples, their performance improved significantly, demonstrating their potential. Our proposed fine-tuning methodology on mBERT still surpassed the few-shot LLM performance, suggesting that for complex, resource-scarce domains like code-switched RE, dedicated fine-tuning with a purpose-built dataset remains highly effective. The use of LoRA [19] also showed competitive results with full fine-tuning, validating its efficiency.

4.3. Ablation Studies

Ablation studies confirmed the importance of key components of our training methodology:

- Code-Switching Awareness Layer: Removing this layer led to a drop of approximately 3-5 F1 points, underscoring its role in discerning language boundaries and integrating linguistic information across languages.

- Adaptive Thresholding and Localized Context Pooling: These techniques, particularly relevant for document-level relations, collectively contributed an improvement of 2-4 F1 points, indicating their effectiveness in handling long-range dependencies and complex inference in code-switched documents [33].

- Data Augmentation: Incorporating synthetic code-switched data [13] led to marginal but consistent gains, especially for low-frequency relation types or language-pair combinations.

These results validate the design choices of our methodology and underscore the necessity of specialized approaches for robust code-switched relation extraction.

5. DISCUSSION

The creation of the MixRED dataset and the development of our tailored training methodology represent a significant step forward in code-switched relation extraction. This work addresses a critical gap in NLP research, providing both a valuable benchmark and an effective solution for processing mix-lingual text.

The MixRED dataset [14] fills a long-standing void, offering a realistic and challenging resource that reflects the complexity of genuine code-switching phenomena [23]. Its size and detailed annotation allow for the development and rigorous evaluation of models specifically designed for this domain. By providing a common ground for evaluation, MixRED will facilitate comparative studies and accelerate advancements in code-switched NLP, much like DocRED [4] and FewRel [16] have done for their respective areas.

Our proposed training methodology demonstrates the efficacy of adapting modern language models to the unique characteristics of code-switching. The substantial performance gains over baselines highlight that generic multilingual models, while a good starting point, benefit significantly from explicit design choices and training strategies that account for inter- and intra-sentential language mixing. The success of techniques like localized context pooling and a code-switching awareness layer underscores the importance of capturing nuanced cross-lingual interactions. Furthermore, the competitive performance with PEFT methods like LoRA [19] indicates that these specialized models can be trained efficiently, making them practical for deployment.

While our model shows strong performance, certain limitations and areas for future work exist.

- Dataset Expansion: While MixRED is large, expanding it to include more language pairs and a wider variety of code-switching styles would further enhance its utility. Real-world code-switching is incredibly diverse, and capturing more of this diversity will be crucial.

- Deeper Linguistic Understanding: Our current methodology primarily relies on surface-level code-switching detection and contextual embeddings. Future work could explore incorporating deeper linguistic features, such as dependency parsing or syntactic structures, that account for grammatical shifts during code-switching, potentially drawing inspiration from natural language generation (NLG) micro-planners [3].

- LLM Integration: While we benchmarked against LLMs, a deeper investigation into how to best fine-tune or prompt-engineer LLMs (e.g., using instruction tuning [22] or more advanced in-context learning strategies [20]) for code-switched RE is warranted. This could involve exploring custom tokenizations for code-switched segments or developing more sophisticated prompting techniques [20, 21, 42] that specifically guide LLMs to handle mix-lingual input more effectively. Approaches similar to ChatGLM [15] which are developed with strong multilingual capabilities might be especially promising.

- Evaluation Metrics: The complexity of code-switching might necessitate more fine-grained evaluation metrics beyond standard F1, particularly for assessing how well models handle different types of code-switching points or language combinations.

- Robustness to Noise: Real-world code-switched data often contains noise, slang, and informal language. Improving the model's robustness to such variations is crucial for practical applications.

Our work lays the groundwork for more advanced research in code-switched NLP, particularly for information extraction tasks. It opens doors for developing more intelligent systems that can seamlessly process the increasingly multilingual and mixed-language content prevalent in digital communication.

6. CONCLUSION

This paper has presented a significant contribution to the field of Natural Language Processing by introducing MixRED [14], a novel and comprehensive dataset for code-switched relation extraction. We have also proposed and empirically validated a robust training methodology tailored to address the unique linguistic challenges of mix-lingual text. Our results demonstrate that by specifically accounting for code-switching phenomena within the model architecture and training process, we can achieve substantial improvements in relation extraction performance on code-switched data. MixRED serves as a crucial resource, enabling future research and benchmarking in this vital area, while our methodology offers a strong foundation for developing more accurate and practical RE systems for the world's increasingly multilingual digital landscape.

REFERENCES

Zhang Y, Zhong V, Chen D, Angeli G, Manning C D. Position-aware attention and supervised data improve slot filling. In Proc. the 2017 Conference on Empirical Methods in Natural Language Processing, Sept. 2017, pp.35–45. DOI: 10.18653/v1/D17-1004.

Zeng X, Zeng D, He S, Liu K, Zhao J. Extracting relational facts by an end-to-end neural model with copy mechanism. In Proc. the 56th Annual Meeting of the Association for Computer Linguistics (Volume 1: Long Papers), Jul. 2018, pp.506–514. DOI: 10.18653/v1/P18-1047.

Gardent C, Shimorina A, Narayan S, Perez-Beltrachini L. Creating training corpora for NLG micro-planners. In Proc. the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2017, pp.179–188. DOI: 10.18653/v1/P17-1017.

Yao Y, Ye D, Li P, Han X, Lin Y, Liu Z, Liu Z, Huang L, Zhou J, Sun M. DocRED: A large-scale document-level relation extraction dataset. In Proc. the 57th Annual Meeting of the Association for Computational Linguistics, Jul. 2019, pp.764–777. DOI: 10.18653/v1/P19-1074.

Luan Y, He L, Ostendorf M, Hajishirzi H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Proc. the 2018 Conference on Empirical Methods in Natural Language Processing, Oct. 31–Nov. 4, 2018, pp.3219–3232. DOI: 10.18653/v1/D18-1360.

Cheng Q, Liu J, Qu X, Zhao J, Liang J, Wang Z, Huai B, Yuan N J, Xiao Y. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In Proc. the Association for Computational Linguistics: ACL-IJCNLP 2021, Aug. 2021, pp.2819–2831. DOI: 10.18653/v1/2021.findings-acl.249.

Zheng S, Wang F, Bao H, Hao Y, Zhou P, Xu B. Joint extraction of entities and relations based on a novel tagging scheme. In Proc. the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2017, pp.1227–1236. DOI: 10.18653/v1/P17-1113.

Wei Z, Su J, Wang Y, Tian Y, Chang Y. A novel cascade binary tagging framework for relational triple extraction. In Proc. the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020, pp.1476–1488. DOI: 10.18653/v1/2020.acl-main.136.

Zhong Z, Chen D. A frustratingly easy approach for entity and relation extraction. In Proc. the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 2021, pp.50–61. DOI: 10.18653/v1/2021.naacl-main.5.

Min B, Jiang Z, Freedman M, Weischedel R. Learning transferable representation for bilingual relation extraction via convolutional neural networks. In Proc. the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Nov. 2017, pp.674–684.

Ni J, Florian R. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In Proc. the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Nov. 2019, pp.399–409. DOI: 10.18653/v1/D19-1038.

Winata G, Aji A F, Yong Z X, Solorio T. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In Proc. the Findings of the Association for Computational Linguistics: ACL 2023, Jul. 2023, pp.2936–2978. DOI: 10.18653/v1/2023.findings-acl.185.

Winata G I, Madotto A, Wu C S, Fung P. Code-switched language models using neural based synthetic data from parallel sentences. In Proc. the 23rd Conference on Computational Natural Language Learning (CoNLL), Nov. 2019, pp.271–280. DOI: 10.18653/v1/K19-1026.

Kong L, Chu Y, Ma Z, Zhang J, He L, Chen J. MixRED: A mix-lingual relation extraction dataset. In Proc. the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), May 2024, pp.11361–11370.

Zeng A, Xu B, Wang B et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv: 2406.12793, 2024. https://arxiv.org/abs/2406.12793, Sept. 2024.

Han X, Zhu H, Yu P, Wang Z, Yao Y, Liu Z, Sun M. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In Proc. the 2018 Conference on Empirical Methods in Natural Language Processing, Oct. 31–Nov. 4, 2018, pp.4803–4809. DOI: 10.18653/v1/D18-1514.

Yang S, Choi M, Cho Y, Choo J. HistRED: A historical document-level relation extraction dataset. In Proc. the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1), Jul. 2023, pp.3207–3224. DOI: 10.18653/v1/2023.acl-long.180.

Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. In Proc. the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1), Aug. 2021, pp.4582–4597. DOI: 10.18653/v1/2021.acl-long.353.

Hu J E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-rank adaptation of large language models. In Proc. the 10th International Conference on Learning Representations, Apr. 2022.

Wan Z, Cheng F, Mao Z, Liu Q, Song H, Li J, Kurohashi S. GPT-RE: In-context learning for relation extraction using large language models. In Proc. the 2023 Conference on Empirical Methods in Natural Language Processing, Dec. 2023, pp.3534–3547. DOI: 10.18653/v1/2023.emnlp-main.214.

Li B, Fang G, Yang Y, Wang Q, Ye W, Zhao W, Zhang S. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. arXiv: 2304.11633, 2023. https://arxiv.org/abs/2304.11633, Sept. 2024.

Li X, Polat F, Groth P. Do instruction-tuned large language models help with relation extraction? In Proc. the 1st Workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd Challenge on Language Models for Knowledge Base Construction (LM-KBC) Co-Located with the 22nd International Semantic Web Conference, Nov. 2023.

Poplack S. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. Linguistics, 1980, 18(7/8): 581–618. DOI: 10.1515/ling.1980.18.7-8.581.

Mihalcea R, Tarau P. TextRank: Bringing order into text. In Proc. the 2004 Conference on Empirical Methods in Natural Language Processing, Jul. 2004, pp.404–411.

Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 1998, 30(1–7): 107–117. DOI: 10.1016/S0169-7552(98)00110-X.

Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proc. the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Nov. 2019, pp.3982–3992. DOI: 10.18653/v1/D19-1410.

Wang Y, Chen M, Zhou W, Cai Y, Liang Y, Liu D, Yang B, Liu J, Hooi B. Should we rely on entity mentions for relation extraction? Debiasing relation extraction with counterfactual analysis. In Proc. the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jul. 2022, pp.3071–3081. DOI: 10.18653/v1/2022.naacl-main.224.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9.

Hendrickx I, Kim S N, Kozareva Z, Nakov P, Séaghdha D Ó, Padó S, Pennacchiotti M, Romano L, Szpakowicz S. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Proc. the 5th International Workshop on Semantic Evaluation, Jul. 2010, pp.33–38.

Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R. The automatic content extraction (ACE) program – Tasks, data, and evaluation. In Proc. the 4th International Conference on Language Resources and Evaluation, May 2004.

Liu L, Li X, He R, Bing L, Joty S, Si L. Enhancing multilingual language model with massive multilingual knowledge triples. In Proc. the 2022 Conference on Empirical Methods in Natural Language Processing, Dec. 2022, pp.6878–6890. DOI: 10.18653/v1/2022.emnlp-main.462.

Nan G, Guo Z, Sekulic I, Lu W. Reasoning with latent structure refinement for document-level relation extraction. In Proc. the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020, pp.1546–1557. DOI: 10.18653/v1/2020.acl-main.141.

Zhou W, Huang K, Ma T, Huang J. Document-level relation extraction with adaptive thresholding and localized context pooling. In Proc. the 35th AAAI Conference on Artificial Intelligence, May 2021, pp.14612–14620. DOI: 10.1609/aaai.v35i16.17717.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. In Proc. the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020, pp.8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

Brown T B, Mann B, Ryder N et al. Language models are few-shot learners. In Proc. the 34th International Conference on Neural Information Processing Systems, Dec. 2020.

Bai J, Bai S, Chu Y et al. Qwen technical report. arXiv: 2309.16609, 2023. https://arxiv.org/abs/2309.16609, Sept. 2024.

Muennighoff N, Wang T, Sutawika L et al. Crosslingual generalization through multitask finetuning. In Proc. the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2023, pp.15991–16111. DOI: 10.18653/v1/2023.acl-long.891.

Yang A, Xiao B, Wang B et al. Baichuan 2: Open large-scale language models. arXiv: 2309.10305, 2023. https://arxiv.org/abs/2309.10305, Sept. 2024.

Touvron H, Martin L, Stone K et al. Llama 2: Open foundation and fine-tuned chat models. arXiv: 2307.09288, 2023. https://arxiv.org/abs/2307.09288, Sept. 2024.