Point and Line: Multilingual Mutual Reinforcement Effect Mix Information Extraction Datasets

Anonymous ACL submission

Abstract

The Mutual Reinforcement Effect (MRE) represents a promising avenue in information extraction and multitasking research. Nevertheless, its applicability has been constrained due to the exclusive availability of MRE mix datasets in Japanese, thereby limiting comprehensive exploration by the global research community. To address this limitation, we introduce a Multilingual MRE mix dataset (MMM) that encompasses 21 sub-datasets in English, Japanese, and Chinese. In this paper, we also propose a method for dataset translation assisted by Large Language Models (LLMs), which significantly reduces the manual annotation time required for dataset construction by leveraging LLMs to translate the original Japanese datasets. Additionally, we have enriched the dataset by incorporating open-domain Named Entity Recognition (NER) and sentence classification tasks. Utilizing this expanded dataset, we developed a unified input-output framework to train an Open-domain Information Extraction Large Language Model (OIELLM). The OIELLM model demonstrates the capability to effectively process novel MMM datasets, exhibiting significant improvements in performance. Furthermore, we conducted a new ablation study to evaluate the MRE across 21 MMM sub-datasets. The results demonstrated that 76% of the datasets exhibited MRE, reinforcing its robustness. Additionally, we applied the MRE datasets to a knowledgeable verbalizer (KV), and the results confirmed that KV constructed by MRE Mix datasets achieved superior KV performance. This further validates the effectiveness of MRE in enhancing IE subtasks.

1 Introduction

Information extraction (IE) Sarawagi et al. (2008) is a significant area of research within natural language processing (NLP). This field has evolved to encompass a variety of subtasks, including sentence classification (Zhang and Wallace, 2015), text classification (Lai et al., 2015), Named Entity Recognition (NER) (Qu et al., 2023; Nadeau and Sekine, 2007; Lample et al., 2016), sentiment analysis (Tan et al., 2023; Medhat et al., 2014; Rodríguez-Ibánez et al., 2023), relationship extraction (Wadhwa et al., 2023; Mintz et al., 2009; Etzioni et al., 2008), and event extraction (Gao et al., 2023; Xiang and Wang, 2019). Traditionally, these IE subtasks have been segregated into distinct categories for processing. In conventional multi-task IE (Sun et al., 2023; Zhao et al., 2020), datasets from various tasks are typically merged and subsequently fine-tuned using a unified model. This process culminates in the extraction of information from multiple subtasks, each directed by task-specific output heads. While this method effectively leverages the internal knowledge of the model across different IE tasks, it does not address the potential interconnections among the tasks themselves. This omission highlights a gap in understanding how these tasks might benefit from exploring their mutual relation-

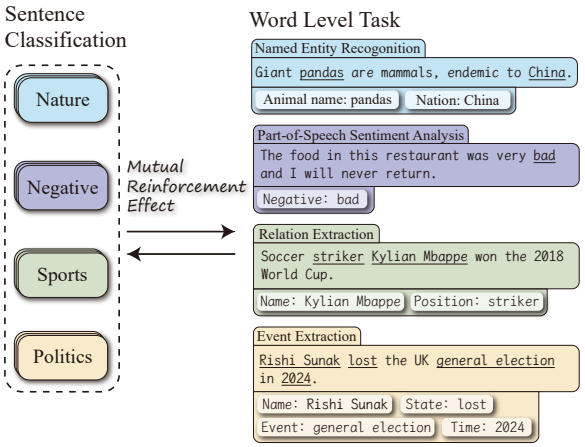

Figure 1: The Mutual Reinforcement Effect between the labels of Word-level labels and text-level label within a same text. **A word-level IE task is a Point, and a text-level IE task is a Line. There is Mutual Reinforcement Effect between the point and the line.**

ships.

The Mutual Reinforcement Effect (MRE) Gan et al. (2023b) introduces a novel approach in multitasking IE, emphasizing task interconnections to enhance performance. MRE categorizes IE subtasks into text-level tasks (e.g., sentence classification, text sentiment analysis) and word-level tasks (e.g., NER). Unlike conventional IE multitasking, which extracts data from various texts, MRE simultaneously performs text-level classification and word-level label-entities pairing within the same text.

MRE categorizes IE tasks into word-level and text-level tasks, analogous to points and lines. Understanding either part helps reinforce the comprehension of the other. Traditionally, IE subtasks have been studied separately, focusing either on points or lines. MRE, however, is the first approach to integrate these two levels, exploring their interdependencies. This not only enhances the performance of IE subtasks but also has implications for future LLM training. When training data is limited, MRE enables dual-level training of LLMs using a single dataset, maximizing its utility and improving model performance.

Figure 1 illustrates MRE in action. The left side depicts sentence classification labels, while the right side shows words with their corresponding labels, representing text-level and word-level tasks, respectively. For example, the sentence 'Giant pandas are mammals, endemic to China.'. is labeled 'nature' and contains entity pairs 'Animal Name: pandas' and 'Nation: China.' This highlights how text-level classification and word-level entity recognition reinforce each other, improving accuracy.

Similarly, in sentiment analysis, a text with many positive words likely conveys a positive sentiment. Conversely, a negative-text classification indicates the presence of negative words. This interaction mirrors human text comprehension, where meaning is derived from individual words and synthesized into an overall context (Gan et al., 2023c).

Figure 2 shows the composition of the Multilingual Mutual Reinforcement Effect Mix (MMM) Datasets, which include seven subdatasets per language across three languages. Notably, SCPOS, focused on sentiment classification and part-of-speech tagging, is larger than others and thus not depicted proportionally. SCNM involves sentence classification and NER, while TCREE covers text classification, relation, and event extraction.

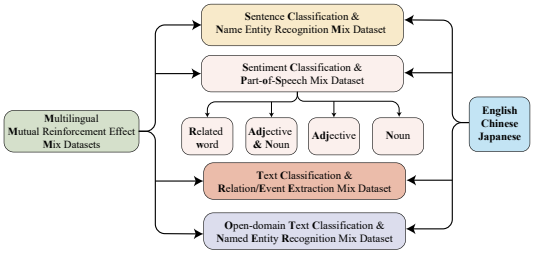

Figure 2: Multilingual Mutual Reinforcement Effect Mix Datasets Names of all sub-datasets.

TCONER leverages an open-domain dataset for text classification and NER.

We translated six MRE mix datasets and expanded the TCONER dataset. To improve LLM performance on IE tasks, we refined the training process by introducing a streamlined input-output scheme, standardizing task handling, and training the LLM with the MMM dataset. The resulting optimized model, OIELLM, outperformed previous models on multiple datasets, demonstrating the effectiveness of using expanded MRE mix datasets. Furthermore, we conducted comprehensive ablation experiments on 21 MMM datasets using an LLM. Notably, 76% of the ablation results demonstrated a positive reinforcement effect, providing strong empirical support for the MRE hypothesis. Additionally, we leveraged word-level information as a Knowledgeable Verbalizer (KV)Hu et al. (2022) to enhance text-level classification tasks. The final experimental results further confirmed the effectiveness of word-level information in improving text-level classification, serving as additional validation for MRE.

Key contributions include:

1. We introduce a framework that minimizes manual annotation by extending the Japanese MRE Mix dataset to English and Chinese and incorporating open-domain text classification and NER tasks. This expansion addresses the lack of open-domain IE subtasks in the original dataset, enhancing its comprehensiveness and applicability.

2. We propose an enhanced Format Converter to train an Open-Domain IE LLM (OIELLM), yielding robust general-purpose IE performance and outperforming conventional methods in MRE mix tasks.

3. A novel ablation experiment method was employed to evaluate the presence of the

MRE across the newly constructed 21 MMM datasets. The empirical results confirm the existence of MRE. Furthermore, by integrating the MRE Mix datasets into the Knowledgeable Verbalizer framework, we indirectly demonstrate that word-level information in MRE significantly enhances performance in text-level classification tasks.

2 Related Work

Datasets. To begin, the MRE mix dataset primarily originates from the SCNM Gan et al. (2023b) dataset in Japanese, followed by the SCPOS (Gan et al., 2023d) and TCREE Gan et al. (2023a) datasets. However, the exclusive use of the Japanese language across these datasets poses significant challenges for researchers attempting to further explore the MRE. Moreover, there has been a growing interest in employing LLMs for dataset construction (Tan et al., 2024; Wadhwa et al., 2023; Li et al., 2023; Laskar et al., 2023). Pioneering studies Huang et al. (2023) have demonstrated the efficacy of LLMs in data annotation, where LLM-annotated datasets have outperformed manually annotated counterparts. For instance, LLMs have been utilized to generate datasets for mathematical problems Lin et al. (2024) and to develop dataset labeling frameworks, such as FreeAL (Xiao et al., 2023a), where the data is initially labeled by LLMs and subsequently refined by smaller models before undergoing a final, more accurate labeling by LLMs again.

These methodologies leverage instructional learning and in-context learning to guide LLMs to respond to specific queries and extract annotated labels from these responses, extract annotated labels, thereby creating a fully labeled dataset. Distinct from previous efforts, the MMM dataset represents the inaugural initiative to translate datasets from lesser-used languages into more widely spoken languages, such as English and Chinese. Furthermore, the newly developed TCONER dataset addresses a critical gap by providing the first open-domain Named Entity Recognition (NER) dataset within the existing framework of the MRE mix dataset.

LLM in Information Extraction. Since the introduction of Pretrained Language Models (PLMs), sequential-to-sequential (seq2seq) based IE models have gained prominence. These developments range from the initial UIE Lu et al. (2022) to later models such as USM Lou et al. (2023) and Mirror (Zhu et al., 2023). All these models are generative in nature, enabling them to handle multiple word-level IE tasks—such as NER, Relation Extraction, and Event Extraction simultaneously. The primary advantage of these generative IE models is their generalizability; they eliminate the need for task-specific fine-tuning across different tasks. Instead, a single model can address all IE subtasks by standardizing the format of inputs and outputs for various tasks. The model is trained across different IE subtasks using these unified formats, aiming to equip a single model with the capability to manage multiple tasks effectively.

With the advent of LLMs, new approaches to IE have emerged, which can be broadly divided into two categories. The first involves direct interaction with LLMs using prompts in a zero-shot or few-shot manner, where the model outputs the desired entities either through multi-round dialog-style prompts or through single-command-based prompts that extract entities in one go (Wang et al., 2023; Wei et al., 2023). The second approach involves fine-tuning LLMs using specialized datasets (Zhou et al., 2023; Xiao et al., 2023b).

Our research distinguishes itself by focusing more intensively on the MRE. We go beyond merely aggregating existing IE sub-datasets for model training. Instead, we develop specialized MRE-enhanced datasets, through which we not only demonstrate but also apply the efficacy of MRE in enhancing information extraction capabilities.

3 Multilingual Mutual Reinforcement Effect Mix Datasets

In this chapter we will explain how to translate MRE mix datasets in small languages into other languages. And how to construct TCONER datasets. And how you can minimize the use of manual labor with guaranteed quality.

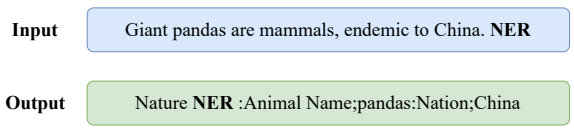

Figure 3: The format of MMM datasets.

3.1 Dataset Translation Framework

First, it is essential to understand the format of the Multilingual Mutual Reinforcement Effect Mix (MMM) dataset. As depicted in Figure 3, the

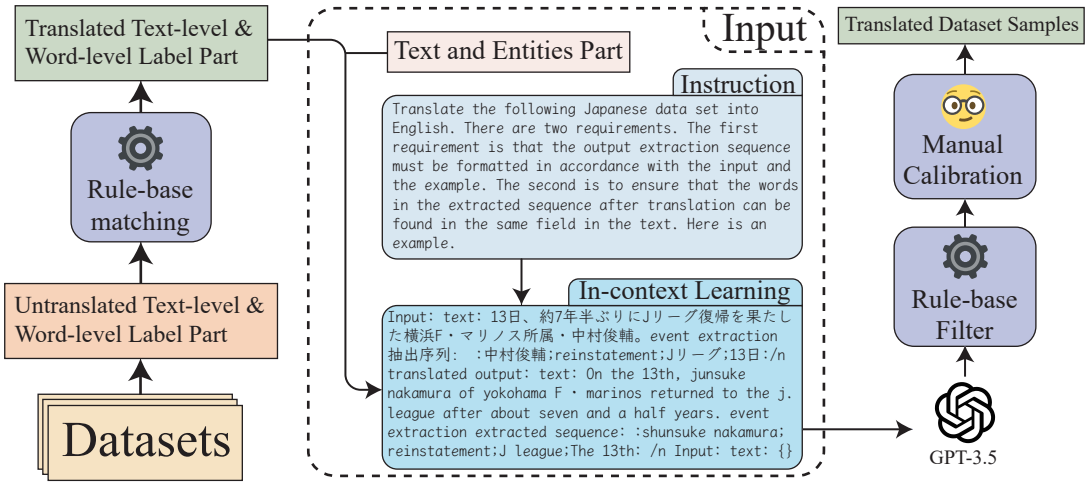

Figure 4: The overview of dataset translation framework.

MMM dataset comprises inputs and outputs. The input section, highlighted in blue, includes both text and a task instruction word, such as "NER." In the output section, shown in green, the initial output is a text-level classification label, followed by the task instruction word "NER". The labeling follows the start and end symbols (i.e., ":", ";") used in the original MRE mixed dataset. This format allows for consistent generation of label-entity pairs regardless of quantity (e.g., ":label1;entities1:label2;entities2..."). Thus, the task instruction word guides the model in producing various word-level extracted information alongside the text-level classification label.

Figure 4 presents a flowchart of the entire dataset translation framework. The process begins on the leftmost side, where six sub-datasets are initially processed using a rule-based matching method, according to their classifications. The labels at both text and word levels are systematically translated into English and Chinese. Given the consistent labeling across datasets, this translation can proceed directly based on predefined rules. For instance, the Japanese label "ポジティブ" is directly translated as "positive." Employing a rule-based approach for label translation is not only quick and precise but also simplifies the subsequent translation of text and entities. Furthermore, these translated labels are input into a LLM along with the untranslated text and entities, serving an auxiliary role in the translation process.

The process involves two main inputs to the LLM, GPT-3.5-Turbo Ouyang et al. (2022): the part with translated labels and the part with untranslated text and entities. We employ both instruction-based and in-context learning (ICL) methodologies for this translation task. As depicted in the central portion of Figure 4, the selection of the instruction template was refined through multiple iterations. Initially, a simple instruction such as "Translate the following Japanese dataset into English." failed to produce satisfactory translations. Consequently, we introduced several constraints to enhance the output quality. These include stipulating that the model's output format must align with the example provided below, with a critical requirement being the accurate translation of entities, ensuring they correspond directly to terms found in the original Japanese text. Additional constraints were applied specifically for Japanese-to-Chinese translations, such as informing the model that labels have been pre-translated and only text and entities require translation. We also instructed the model to ensure comprehensive translation into Chinese. Furthermore, a one-shot example of ICL was provided to demonstrate the desired outcome, guiding the model to generate translations strictly adhering to the specified format.

Finally, we obtained the translated dataset. However, due to the inherent unpredictability of LLM outputs, it is not always guaranteed that the outputs will conform to the expected format, even when the inputs are consistent. To address this, we implemented a dual-component rule-based filtering mechanism. The first component involves removing samples containing any residual Japanese characters from the translated data. The second component entails verifying whether the translated entities exactly match words in the text. Samples that do not meet this criterion are excluded. Additionally,

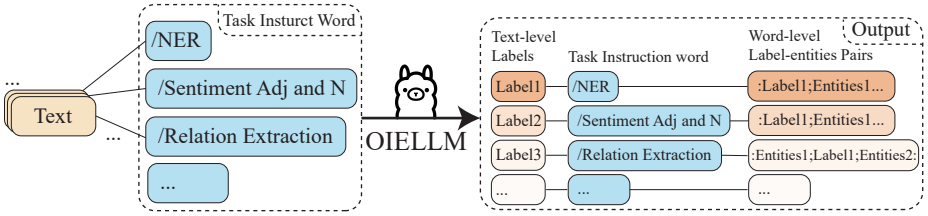

Figure 5: The input and output of Open-domain Information Extraction Large Language Model (OIELLM).

this step assesses whether the pairings of labels and entities adhere to the formatting standards of the MMM dataset.

Despite the substantial reduction in dataset size resulting from the first two steps—translation and filtering—the remaining data exhibit exceptionally high translation quality. The final dataset undergoes a manual review and correction process, which ensures maximum accuracy while minimizing the reliance on manual labor. We enlisted ten graduate students proficient in two of the three languages—Chinese, English, and Japanese—to conduct the final round of data validation. To ensure the accurate recognition of rare or specialized terms, we instructed them to consult authoritative dictionaries such as the Oxford Dictionary for verification and refinement. This approach outlines our tailored dataset translation framework, designed to accommodate the specific characteristics of the MMM dataset. With minimal modifications, this framework can be adapted for translating datasets for other tasks, effectively addressing the scarcity of datasets in lesser-used languages. And construction results details of MMM dataset can find in Appendix C.

4 Open-domain Information Extraction Large Language Model

In this chapter, we outline methodologies to enhance the performance of existing models and techniques for processing MRE mix datasets, aiming to surpass previous benchmarks. Before delving into the specifics of the Open-domain Information Extraction Large Language Model (OIELLM), it is imperative to justify the necessity for a distinct model tailored to MMM datasets.

Firstly, MRE mix datasets differ significantly from traditional IE tasks as they require simultaneous output of text-level labels and word-level label-entity pairs. Consequently, standard sequence labeling models are inadequate for handling these demands directly. Furthermore, existing generative IE models and methodologies have solely focused on producing word-level label-entities, neglecting text-level labels altogether.

The primary objective of MRE mix datasets is to investigate the interplay between text-level and word-level annotations. By leveraging this synergistic relationship, we aim to concurrently enhance the performance of both tasks. This model improves textual understanding by learning both tasks in tandem. Additionally, the MRE framework can contribute to model interpretability, drawing inspiration from cognitive processes that mimic human reasoning.

This study introduces a specialized model for the MMM dataset and examines whether MRE improves various IE subtasks in LLMs. Instead of using QA-style dialogues, we follow earlier generative IE work that relies on a generic framework. Thus, we adopt a tailored input-output scheme for the MMM dataset, departing from traditional dialogue-based methods.

Figure 5 illustrates the input and output formats of our enhanced OIELLM. The fundamental unit of analysis in both input and output is words, reflecting our understanding of the tokenization principle utilized by LLMs, which typically focuses on words or phrases. By omitting the dialog prompt, we do not compromise the LLM's comprehension of the task. This adjustment not only reduces the input-output length but also simplifies the LLM's processing, thereby enhancing operational speed.

Each text processed is prefixed with task-specific instruction words, which define the task type and guide the model's subsequent output generation. In our format, all task instruction words in the input are introduced by a special symbol "/", which serves to delineate the task words from the main text. This separation is crucial for distinguishing between text-level labels and word-level label-entity pairs in the output.

The combined text and task instruction words are then fed into the OIELLM, with the output comprising both text-level labels and word-level label-

Japanese		SCNM			SCPOS: RW			SCPOS: Adj & N	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	42.07	7.54	1.97	57.20	0	0	28.97	5.97	0
GPT-4o-mini	0.27	20.61	0				1.33	3.01	0
USA-7B	-	-	-	53.27	40.80	7.67	91.33	**81.68**	**9.63**
GIELLM-13B-jp	85.47	84.46	54.2	86.01	**66.61**	**17.39**	93.23	47.35	0.20
OIELLM-8B	84.73	88.53	61.93	86.50	54.76	12.40	89.13	14.88	0.40
OIELLM-8B*	87.30	**89.28**	**64.00**	88.20	53.79	12.30	89.63	15.84	0.73
OIELLM-13B	**89.00**	86.33	57.70	**94.60**	52.36	11.90	**95.20**	11.94	0.20
Japanese		SCPOS: Adj			SCPOS: N			TCREE	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	65.50	0.31	0.87	39.60	6.79	0	57.20	0	0
GPT-4o-mini	0.03	0.18	0	0	2.94	0	0	0	0
USA-7B	91.43	45.51	51.77	92.03	**81.30**	**9.73**	-	-	-
GIELLM-13B-jp	93.67	45.06	**55.67**	92.83	46.42	0.33	**97.47**	**79.01**	77.89
OIELLM-8B	87.13	74.96	53.07	87.77	22.92	0.50	95.07	74.92	83.69
OIELLM-8B*	89.93	**75.33**	54.93	90.63	23.69	0.63	96.98	74.42	**84.19**
OIELLM-13B	**94.00**	60.69	42.50	**94.70**	18.07	0.60	97.08	73.82	84.19

English		SCNM			SCPOS: RW			SCPOS: Adj & N	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	53.50	0.04	0	14.78	2.11	0.12	68.63	13.62	0.33
GPT-4o-mini	0	0.03	0	0	0.04	0	0	0	0
OIELLM-8B	82.30	81.36	52.53	72.17	49.60	11.82	76.57	18.00	1.67
OIELLM-8B*	**85.43**	**82.38**	**55.43**	**74.75**	**49.93**	**12.81**	79.77	**19.28**	2.27
OIELLM-13B	84.80	80.68	50.60	95.07	46.64	12.19	**94.30**	18.59	**3.20**
English		SCPOS: Adj			SCPOS: N			TCREE	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	6.97	0.26	0.03	0.53	0.08	0	12.87	0	0
GPT-4o-mini	0	0	0	0	0	0	0	0	0
OIELLM-8B	75.47	51.85	32.33	76.10	28.67	1.27	80.87	21.77	33.67
OIELLM-8B*	76.60	**51.95**	33.17	78.67	27.45	**0.73**	80.23	**25.90**	22.37
OIELLM-13B	**94.40**	50.56	**38.40**	**95.30**	**28.36**	0.60	**89.90**	23.50	**22.60**

Chinese		SCNM			SCPOS: RW			SCPOS: Adj & N	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	41.63	9.57	2.30	50.77	2.08	0.78	59.33	7.18	0.40
GPT-4o-mini	5.20	18.52	0.50	12.14	7.49	0.11	0.53	1.36	0
OIELLM-8B	84.90	**71.90**	46.40	89.29	45.75	9.93	92.33	8.75	0.33
OIELLM-8B*	86.33	69.97	**46.77**	92.27	**46.20**	**10.60**	94.50	**8.46**	0.40
OIELLM-13B	**87.70**	68.12	41.60	**95.03**	43.32	8.72	**94.90**	8.42	**0.50**
Chinese		SCPOS: Adj			SCPOS: N			TCREE	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	56.27	0.19	0.07	53.07	3.11	0.53	59.33	7.18	0.40
GPT-4o-mini	27.37	1.43	0.20	5.33	1.36	0	0	0	0
OIELLM-8B	93.73	60.96	53.00	92.63	28.32	0.63	91.73	58.12	56.41
OIELLM-8B*	95.80	64.51	**57.63**	94.97	**28.91**	**1.30**	95.06	**59.54**	**58.83**
OIELLM-13B	**96.00**	60.68	54.90	**95.20**	27.77	1.00	**95.26**	56.91	56.00

TCONER		English			Japanese			Chinese	
Model	TL	WL	ALL	TL	WL	ALL	TL	WL	ALL
GPT-3.5-Turbo	23.87	4.78	0	23.87	2.24	0.17	29.47	2.97	**0.57**
GPT-4o-mini	2.93	4.06	0	0	3.68	0	0.03	6.12	0
OIELLM-8B	24.80	21.12	0.20	27.70	13.83	**0.20**	33.73	**18.87**	0
OIELLM-8B*	37.13	**23.05**	**0.30**	41.40	**14.24**	0.17	**48.27**	18.06	0.17
OIELLM-13B	**40.30**	19.23	**0.30**	**43.40**	13.02	0	47.70	15.72	0.30

Table 1: The F1 score of MMM datasets. TL F1 score：Text-Level Classification task(e.g. Sentence/Text Classification). WL F1 score: Word-level Label-Entities pairs task(e.g. NER, RE, EE etc.). ALL F1 score: TL and WL are correct simultaneously in one sentence. Note:

entity pairs. Our labeling convention adheres to the format used in the previous MRE mix datasets, utilizing ":" and ";" to ensure consistency and clarity.

In summary, by standardizing the input and output structures and clearly defining task instruction words, our modified OIELLM effectively processes all sub-datasets within the MMM framework.

5 Experiment

In this chapter, we provide a comprehensive overview of our experimental setup, including dataset construction, training procedures, and evaluation metrics.

5.1 Details of OIELLM Training

We began by selecting baselines: USA-7B (IL + ICL)[1] and GIELLM-13B-jp[2], previously utilized for processing the MRE mixed datasets, served as comparative models. For the foundational architecture of OIELLM, we chose the latest Instruct and Base version of LLaMA3-8B[3]. Since LLaMA3 does not offer a 13B version, we incorporated the LLaMA2-13B Touvron et al. (2023) model as well.

We attempted to evaluate the MMM dataset using the GPT-3.5-Turbo model and GPT-4o-mini (1-shot with In-context and Instruction Learning); however, this model failed to produce the expected information and was unable to maintain a consistent format, despite being provided with an adequate number of few-shot examples for training. The resulting F1-score was near zero. Consequently, we decided not to select the GPT-3.5-Turbo model for further testing in our study.

OIELLM was fine-tuned using full parameters based on these three models. Training was conducted at BF16 precision, while inference was performed at FP16. The training spanned 3 epochs with a learning rate of 1e-5, utilizing computational resources including three A800 80GB and three RTX 6000 Ada 48GB GPUs, with training durations ranging from 12 to 20 hours. For the training and test sets, Comprehensive statistics on the training and test sets are available in Appendix Table 6, 7.

5.2 Evaluation

We employed the F1 score as our primary metric for evaluation. Initially, the model's output was bifurcated into two segments based on the task-specific instruct word: the Text-level Label and the Label-entities pairs. Subsequently, Label-entities pairs were delimited using start-end symbols (i.e., ":", ";"). Each Label-entity pair was treated as an individual element within the set. The F1 score was segmented into three categories: Text-level (TL), Word-level (WL), and ALL. These represent the F1 scores at respective levels and the aggregate F1 score when both levels are accurately predicted in an output. For detailed methodologies, including codes and formulas, please refer to Appendix E.

[1] https://huggingface.co/ganchengguang/USA-7B-instruction-incontext-learning

[2] https://huggingface.co/ganchengguang/GIELLM-13B-jpllm

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

6 Results

Table 1 presents the experimental results of three OIELLM models trained on 21 MMM sub-datasets. Notably, the model designated with an asterisk, OIELLM-8B, was trained using the LLaMA3-8B-Instruct framework, whereas the remaining models were based on the LLaMA3-8B-Base framework. These results demonstrate the enhanced performance of OIELLM in handling Japanese data after incorporating multilingual capabilities. Impressively, OIELLM's performance surpassed that of GIELLM-13B-jp on half of the datasets, despite GIELLM-13B-jp being a model specifically tailored for Japanese. This observation supports the hypothesis that integrating multilingualism and multitasking can more effectively leverage the knowledge embedded in the pre-training of multilingual LLMs.

However, OIELLM's performance on the TCONER task was suboptimal, which we attribute to insufficient training data. Given that open-domain tasks require extensive and diverse datasets compared to domain-specific tasks, the limited data may have hindered the model's performance. This area will be a focus of our future research, aiming to understand and improve the data dependencies of OIELLM in open-domain contexts. Due to the high cost of accessing GPT-4o, we conducted experiments on MMM datasets using GPT-3.5-Turbo and GPT-4o-mini only. The low F1 scores of the GPT series models can be attributed to two key factors. First, we impose strict constraints on the output format—any deviation, even a single incorrect symbol, is considered an error. Compared to previous evaluations based on accuracy, our exclusive use of the F1 score in this experiment further contributes to the lower results. Second, the GPT series models have not undergone supervised fine-tuning (SFT) specifically for MRE, making it particularly challenging for them to perform both text-level and word-level tasks simultaneously on the same input. This limitation underscores the necessity of training dedicated IE LLMs optimized for MRE, highlighting their critical role in achieving superior performance.

7 Ablation Experiment of MMM Datasets

The detailed of ablation experiments, including their setup and configuration, in Appendix A and B.1. From the results in Table 2, we observe that for the first six fixed-label datasets, models trained

English	SCNM	SCPOS:RW	SCPOS:adj&n	SCPOS:adj	SCPOS:n	TCREE
w/o TLI	80.97	48.79	**33.29**	56.04	**28.79**	16.43
with TLI	**81.28**	**48.99**	32.42	**56.75**	27.71	**18.43**
w/o WLI	82.40	72.41	77.27	73.73	77.07	82.23
with WLI	**83.90**	**73.15**	**77.60**	**75.70**	**77.73**	**83.33**
Chinese	**SCNM**	**SCPOS:RW**	**SCPOS:adj&n**	**SCPOS:adj**	**SCPOS:n**	**TCREE**
w/o TLI	**73.35**	**44.36**	28.67	9.68	29.06	55.10
with TLI	72.81	43.30	**29.17**	**9.73**	**29.34**	**56.31**
w/o WLI	83.17	89.07	91.03	**93.67**	91.80	93.64
with WLI	**83.93**	**90.95**	**92.37**	92.07	**93.63**	**94.85**
Japanese	**SCNM**	**SCPOS:RW**	**SCPOS:adj&n**	**SCPOS:adj**	**SCPOS:n**	**TCREE**
w/o TLI	87.92	69.47	63.80	50.70	**67.23**	80.87
with TLI	**88.22**	**69.92**	**63.89**	**51.03**	66.24	**81.37**
w/o WLI	83.60	87.10	88.13	87.93	88.37	**94.86**
with WLI	**85.87**	**89.50**	**89.17**	**89.90**	**90.57**	94.46
TCONER	English		Chinese		Japanese	
w/o TLI	**20.22**		17.28		13.19	
with TLI	19.85		**17.82**		**13.39**	
w/o WLI	**36.50**		**44.07**		38.97	
with WLI	35.53		43.33		**43.30**	

Table 2: The results of text-level information (TLI) and word-level information (WLI) comparison experiments.

with the inclusion of additional information consistently outperform those trained without it. **76% of the experimental results demonstrated that the inclusion of one level of information would have a facilitating effect on another level of information**. These findings strongly support the MRE hypothesis, demonstrating that mutual reinforcement exists between word-level and text-level classification tasks. A well-balanced combination of both classification levels enhances the LLMs ability to understand and perform across tasks. Specifically, comprehension of one task level (e.g., text-level) facilitates and strengthens the understanding of the other (e.g., word-level).

This insight not only advances our understanding of how LLMs tackle natural language tasks but also reflects a broader principle underlying human cognition: the mutual reinforcement between different levels of text comprehension mirrors how humans naturally process and understand language.

As illustrated by the results of the open-domain text classification and NER tasks at the bottom of Table 2, approximately half of the outcomes do not surpass those achieved by the model trained without Level Information. We attribute this to the nature of certain open-domain datasets, which contain multiple labels; in such cases, not all WLI contributes positively to TLI. The presence of these uncorrelated WLIs and TLIs leads to a decline in overall performance. However, in the Chinese and Japanese TCONER datasets, we observe improved results after incorporating Level Information. This improvement suggests that the MRE is more effective in languages based on Chinese characters, in contrast to those that use alphabetic writing systems, such as English.

8 Conclusion and Future Work

In this work, we propose an auxiliary framework for automated dataset translation, eliminating dataset scarcity as a barrier to low-resource language research. Additionally, we construct the TCONER dataset, addressing the absence of open-domain IE tasks in the MRE Mix datasets. By training OIELLM on the newly developed MMM dataset, we further validate the effectiveness of our approach. Finally, through ablation experiments, we empirically verify the MRE hypothesis. Moreover, we apply the MMM dataset to KV tasks, achieving promising results.

9 Limitations

Due to resource constraints, we were unable to employ the higher-performing GPT-4o OpenAI (2023) model as the base for our dataset translation framework. Consequently, this model was

also not utilized during the testing phase on the dataset. In future work, we aim to leverage a more advanced model, such as the GPT-4o, to evaluate the MMM dataset, provided that the necessary resources become available. It is important to note that the dataset translation framework proposed in this study is not designed to fully replace human translators. Instead, it leverages LLMs to reduce the time and effort required for translating and processing simpler examples, allowing human expertise to be allocated to more complex and nuanced cases. Ultimately, human verification remains essential to ensure the accuracy and quality of all translated results. Therefore, we do not explicitly evaluate the quality of the translated datasets, as all translations ultimately require human verification and refinement.

References

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023a. Giellm: Japanese general information extraction large language model utilizing mutual reinforcement effect. *arXiv preprint arXiv:2311.06838*.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023b. Sentence-to-label generation framework for multi-task learning of japanese sentence classification and named entity recognition. In *International Conference on Applications of Natural Language to Information Systems*, pages 257–270. Springer.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023c. Think from words (tfw): Initiating human-like cognition in large language models through think from words for japanese text-level classification. *arXiv preprint arXiv:2312.03458*.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023d. Usa: Universal sentiment analysis model & construction of japanese sentiment text classification and part of speech dataset. *Preprint*, arXiv:2309.03787.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Md Tahmid Rahman Laskar, Mizanur Rahman, Israt Jahan, Enamul Hoque, and Jimmy Huang. 2023. Can large language models fix data annotation errors? an empirical study using debatepedia for query-focused text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10245–10255.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Qingwen Lin, Boyan Xu, Zhengting Huang, and Ruichu Cai. 2024. From large to tiny: Distilling and refining mathematical expertise for math word problems with weakly supervision. *arXiv preprint arXiv:2403.14390*.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. *Preprint*, arXiv:2301.03282.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

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.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. 2023. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862.

Sunita Sarawagi et al. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2023. Learning implicit and explicit multi-task interactions for information extraction. *ACM Transactions on Information Systems*, 41(2):1–29.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023a. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023b. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *arXiv preprint arXiv:2312.15548*.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3239–3248.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Tong Zhu, Junfei Ren, Zijian Yu, Mengsong Wu, Guoliang Zhang, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2023. Mirror: A universal framework for various information extraction tasks. *arXiv preprint*. ArXiv:2311.05419 [cs].

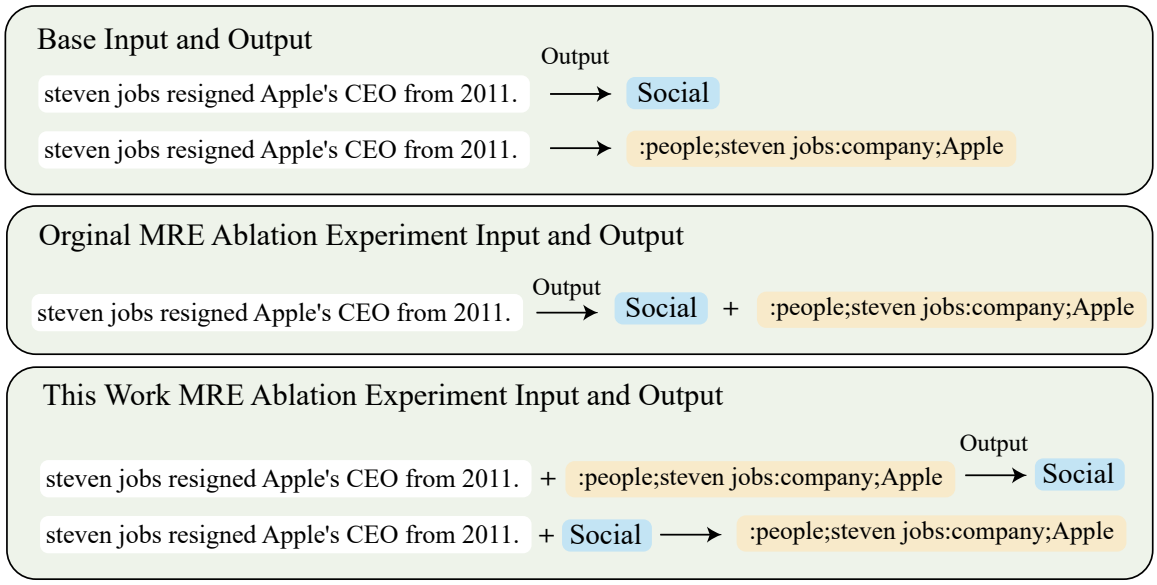

Figure 6: The figure shows the inputs and outputs of the traditional ablation experiment for the MRE task and the new empirical MRE experiment proposed in this work.

A Empirical Experiment of Mutual Reinforcement Effect

The three format of fine-tuned language models used for ablation experiments are shown in Figure 6. The sentence on the left represents the input, with the plus sign indicating the addition of Word-level Information (WLI. i.e. Word-level Task) or Text-level Information (TLI. i.e. Text-level Task), which are appended to the sentence to form the full input. The arrows represent the output produced by language model. The distinctions between the models are clearly illustrated.

First, the top model in Figure 6 shows the input-output format for the traditional IE task, where language models are fine-tuned on a basic input sentence. The model then outputs either classified labels or extracted label-entity pairs. This approach treats the two tasks—word-level label extraction and text-level classification—independently, with no shared information between them.

In contrast, the middle section of Figure 6 illustrates the input-output format for the original MRE task. While the input remains a single sentence, the model is expected to output both word-level label-entity pairs and text-level classification labels simultaneously. Thus, during MRE fine-tuning, the model learns to capture both levels of information, integrating the two tasks.

Finally, the bottom section of Figure 6 presents the input-output format of our proposed ablation experiment designed to validate MRE. Unlike the previous two formats, this approach aims to verify the existence of shared knowledge between word-level and text-level tasks. Specifically, we introduce WLI and TLI to both levels of tasks to assess whether enhancing one task also improves the other. For example, by adding word-level label-entity pairs to the input text and asking the model to output the text-level classification label, we can evaluate whether the additional word-level information assists in text classification. Similarly, if adding text-level information to the input improves the extraction of word-level label-entity pairs, it suggests the presence of an MRE between the two tasks.

As showed in Figure 7, the LLM is fine-tuned with all parameters using revised input and output formats. The input sequence is directly concatenated with either WLI or TLI, while the output consists solely of TLI or WLI. No additional instruction templates or prompt words were incorporated in this process. We deliberately concatenated the text with WLI or TLI without extra modifications to minimize the potential influence of extraneous words or sentences on the model's output, which could affect the accuracy of our comparative experiments. By using only this basic spliced input and raw output, we aim to investigate whether tasks at one level facilitate tasks at another, while controlling for other confounding factors.

To test this hypothesis, we conducted ablation experiments on 21 sub-datasets of Multilingual MRE

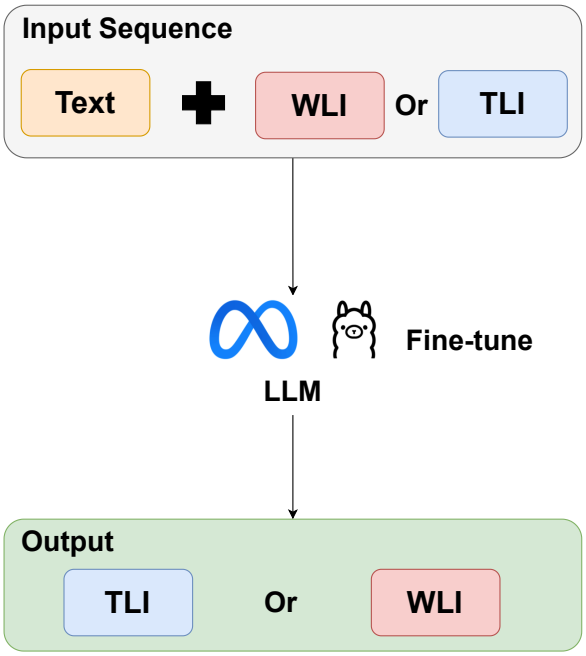

Figure 7: The figure illustrates the flow of an empirical MRE experiment using the new approach.

Mix (MMM) datasets. The results were analyzed to further deepen our understanding of MRE and its implications.

B Word-level Information as Knowledgeable Verbalizer

To enhance the application of the MRE approach in real-world contexts, we have selected the few-shot learning task for text classification as our experimental setup. In MRE, word-level information plays a crucial role in text-level classification. Hence, we utilize the high-frequency words from word-level information as knowledgeable verbalizers (KV) Hu et al. (2022) to examine their impact on the performance of the text classification task.

The entire process of prompt learning is illustrated in Figure 8. Initially, a target classification text is provided, followed by the inclusion of a prompt template to guide the model in predicting the label at the designated mask position. Our sample dataset comprises five labels. We employ the top 100 words from the word-level information as the knowledgeable verbalizer, meaning that each of the five categories has 100 high-frequency words selected from the word-level information. When calculating the actual probability of a label, the model computes the probability of all these 500 words and then aggregates the total probability based on the respective broad classification. Ultimately, we obtain five probabilities that integrate the individual verbalizers. The label with the highest probability is chosen as the final predicted label.

In conclusion, this outlines the detailed principle behind the KV. In the original experimental setup, label-related high-frequency words were sourced directly from a relation word search website, where commonly used vocabulary was analyzed to identify relevant terms. While these words may be highly pertinent across a wide range of web texts, not all of them are necessarily associated with the labels of a specific dataset. As a result, some of these words may not only fail to enhance label prediction but could potentially introduce negative effects. This highlights the suitability of the WLI component from the MRE-mixed dataset as a replacement for the KV. Furthermore, if the performance of the WLI-based KV surpasses that of the original baseline KV, it would support the argument that WLI contributes positively to label prediction in text classification tasks. This, in turn, would verify the presence of the MRE.

B.1 Experiment Setup of Ablation and KV Experiment

For the empirical experiments on fine-tuning, we selected the LLaMA3-8B[4] model[5] as the base model to perform a series of fine-tuning and inference tasks. We opted not to use the LLaMA3-8B-Instruct version because it is more tailored for question-answering tasks, with prompts structured as instructions. Through a comparative analysis of LLaMA3-8B and its instruct-tuned counterpart, we observed that the base LLaMA3-8B model achieved better performance on fundamental IE tasks. Therefore, we decided to use LLaMA3-8B as the foundation for our experiments.

For the WLI as KV application comparison experiments, we employed the T5-base Raffel et al. (2020) model as the base model. Specifically, for the English portion of the MMM dataset, we used the original Google T5-base[6]. For the Chinese section, we selected the Mengzi-T5-base[7], which is optimized for Chinese tasks. Lastly, for the Japanese part of the MMM dataset, we utilized T5-base-Japanese[8].

For the fine-tuning experiment, the entire train-

[4]https://ai.meta.com/blog/meta-llama-3/
[5]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[6]https://huggingface.co/google-t5/t5-base
[7]https://huggingface.co/Langboat/mengzi-t5-base
[8]https://huggingface.co/sonoisa/t5-base-japanese

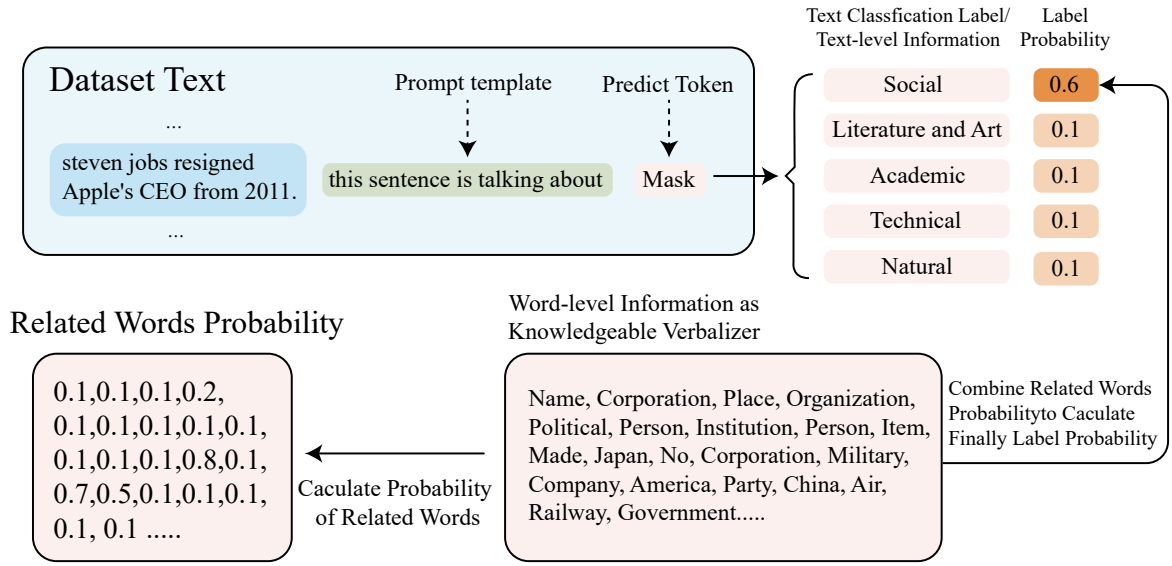

Figure 8: The figure demonstrates how word-level information is utilized as a Knowledgeable Verbalizer to assist in text-level classification tasks. Additionally, it provides a detailed explanation of the functioning of the Knowledgeable Verbalizer.

Datasets	**Text-level**	**Word-level**
SCNM	Society, Literature, Academia, Technology, Nature	people, corporations, political organizations, other organizations, places, facilities, products, and events
SCPOS:RW	positive, negative	positive, neutral, negative
SCPOS:N	positive, negative	positive, neutral, negative
SCPOS:Adj	positive, negative	positive, negative
SCPOS:N & Adj	positive, negative	positive, neutral, negative
TCREE	sports, film, women, IT, advertising	affiliation, occupation, starring, director, age, product, goods, performances, wins, broadcasts, public appearances, launches, retirements
TCONER	Entertainment, Politics Medical, Health, education Tech, Healthcare, News finance, Biolog, etc.	date, location, organization Title, Person, City Law, Number, Concept TV Show, Object, etc.

Table 3: The table presents seven distinct types of MRE mixed datasets, each available in Chinese, English, and Japanese, resulting in a total of 21 sub-datasets. Among them, the TCONER dataset corresponds to an open-domain dataset, where only a subset of the labels is provided, rather than a comprehensive list of all possible labels. (SCNM: Sentence Classification and Named Entity Recognition Mix Dataset. SCPOS: Sentiment Classification and Part-of-Speech Dataset. RW: Relation Word. N: Noun. Adj: Adjective. N & Adj: Nouns and Adjective. TCREE: Text Classification and Relation & Event Extraction Dataset. TCONER: Open-domain Text Classification and NER mix dataset)

ing set was utilized to fully parameterize the fine-tuned LLMs. Subsequently, 1,000 samples were randomly selected from the test set three times, and the results from these three trials were averaged to produce the final performance score. The evaluation metric employed was the F1 score.

The hyperparameters for training were configured as follows: the number of training epochs was set to 3, and the learning rate was initialized at 1e-5. The AdamW optimizer was used, with 100 warm-up steps. Training was conducted on three RTX A6000 Ada GPUs, each with 48 GB of memory. To optimize GPU memory usage, BF16 precision was applied during training, and FP16 precision

was employed for inference.

Second, for the experiments involving the knowledgeable verbalizer, we utilized the OpenPromptDing et al. (2021)[9] framework to efficiently set up the experimental environment. All datasets were divided into training and test sets. From the training set, we randomly selected 20 samples per category, based on the label types, to form the prompt experiment's training subset. Each experiment was trained for 2 epochs, with all other hyperparameters—such as the learning rate—kept consistent across experiments. The only variation lay in the construction method of the KV.

For the KVs based on the original approach, we leveraged ChatGPT-4o[10] to generate the top 100 most relevant words for each label. In contrast, for KVs constructed using the WLI-based method, we developed a custom processing script. The script segmented all words from the WLI section of each dataset, identified high-frequency terms, and used them to construct the WLI-based KVs.

B.2 Results of Word-level Information as Knowledgeable Verbalizer

The next result involves the use of WLI as the relevant word for constructing KVs. We compare the performance of KVs constructed using the original method with those built using WLI in a text classification task. Since KV construction requires a fixed label structure, the open-domain TCONER dataset, which has an unfixed label schema, was excluded from this experiment.

As shown in Table 4, across 18 sub-datasets in English, Chinese, and Japanese, the WLI-based KVs achieved the highest performance in 16 datasets. Moreover, for most sentiment classification datasets, KVs constructed with WLI significantly outperformed those generated by the original method in terms of F1 scores. These results not only demonstrate the effectiveness of WLI in enhancing general text classification tasks but also highlight its particular value in sentiment classification. This is likely because sentiment classification heavily relies on correctly identifying the sentiment polarity of individual words within the text, which aligns with WLI's strengths.

[9]https://github.com/thunlp/OpenPrompt

[10]https://chatgpt.com/

C Construction of TCONER

In the original MRE mix datasets, relation and event extraction tasks are open-domain, implying that the labels are not predefined. However, the label set is limited to only a dozen options. Given this context, we constructed a new dataset, termed TCONER, based on an open-domain Named Entity Recognition (NER) dataset[11] (Zhou et al., 2023). The labels at the text level in the TCONER dataset are also open-domain. To annotate this dataset, we initially employed the GPT-3.5-Turbo model to assign open-domain text-level labels. Subsequent manual verification and annotation were conducted to ensure accuracy and consistency, resulting in the finalized TCONER dataset. Similarly, we translated the constructed English TCONER dataset using the dataset translation framework. The TCONER dataset was translated into Japanese and Chinese.

Table 5 presents the statistics of the final translation results. Due to the high costs associated with the use of a premium API, we limited our study to 10,000 samples from each of three sub-datasets within SCPOS and the TCONER dataset, which contains 180,000 entries. These 10,000 samples, retained post-translation, proved to be an ample test set. It was observed that there was a greater data loss when translating into Chinese compared to English. This discrepancy may be attributed to the training data predominance of English in OpenAI's GPT-3.5-Turbo model, resulting in superior performance in English-related tasks. For instance, in the SCNM and TCREE datasets, the Japanese to English translation accuracy exceeded 80%. Conversely, the translation results from English to Chinese in the TCONER dataset were markedly better than those from English to Japanese. This further confirms that GPT-3.5-Turbo exhibits enhanced effectiveness with major languages compared to lesser-used ones.

D Statistical Results of Train and Test Dataset in OIELLM

As shown in Tables 6 and 7, the statistics for the complete training and test sets of the MMM dataset. The MMM dataset was segmented into 21 sub-datasets. Training set sizes were assigned based on the sizes of these sub-datasets, categorized into

[11]https://huggingface.co/datasets/Universal-NER/Pile-NER-type?row=0

English	SCNM	SCPOS:RW	SCPOS:adj&n	SCPOS:adj	SCPOS:n	TCREE
Origin KV	62.95	80.42	80.40	78.87	81.95	**86.52**
WLI KV	**63.24**	**83.99**	**87.40**	**87.37**	**88.70**	85.82
Chinese	**SCNM**	**SCPOS:RW**	**SCPOS:adj&n**	**SCPOS:adj**	**SCPOS:n**	**TCREE**
Origin KV	67.38	78.37	**91.90**	84.48	84.45	93.04
WLI KV	**71.96**	**87.97**	82.92	**88.38**	**87.23**	**93.95**
Japanese	**SCNM**	**SCPOS:RW**	**SCPOS:adj&n**	**SCPOS:adj**	**SCPOS:n**	**TCREE**
Origin KV	73.26	30.20	67.23	73.71	73.71	73.11
WLI KV	**73.91**	**52.90**	**81.74**	**85.67**	**88.31**	**77.24**

Table 4: The results of word-level information (WLI) as knowledgeable verbalizer experiments. Compare with original KV construction method. Evaluation task is text classification task.

Dataset	SCNM	SCPOS: RW RW	SCPOS: Adj & N
Japanese	**5343**	**2000**	**187528**
English	4449	1312	4801
Chinese	3177	1406	3937

Dataset	SCPOS: Adj	SCPOS: N	TCREE
Japanese	**187528**	**187528**	**2000**
English	9132	5027	1910
Chinese	7413	3920	1491

Language	English	Japanese	Chinese
TCONER	**45888**	6791	9047

Table 5: Statistical results of the translated MMM dataset. (Due to resource constraints, we extracted only 10,000 samples as translation objects from each of the three SCPOS sub-datasets and the TCONER dataset.)

Dataset	SCNM	SCPOS: RW RW	SCPOS: Adj & N
Japanese	1000	1000	1000
English	1000	500	1000
Chinese	1000	500	1000

Dataset	SCPOS: Adj	SCPOS: N	TCREE
Japanese	1000	1000	1000
English	1000	1000	500
Chinese	1000	1000	500

Language	English	Japanese	Chinese
TCONER	2000	2000	2000

Table 6: Statistical results of train sets of OIELLM.

three groups: 500, 1000, and 2000 samples. Samples beyond these numbers were allocated to the test sets.

Dataset	SCNM	SCPOS: RW RW	SCPOS: Adj & N
Japanese	4343	1000	186528
English	3449	812	3801
Chinese	2177	906	2937

Dataset	SCPOS: Adj	SCPOS: N	TCREE
Japanese	186528	186528	1000
English	8132	4027	1410
Chinese	6413	2920	991

Language	English	Japanese	Chinese
TCONER	43888	4791	7047

Table 7: Statistical results of test sets.

E Calculate Detail of F1 Score

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

$$\text{precision} = \frac{|Real \cap Generated|}{|Generated|} \quad (2)$$

$$\text{recall} = \frac{|Real \cap Generated|}{|Real|} \quad (3)$$

F Case Study of Input and Output Format with OIELLM in MRE mix datasets

Algorithm 1 Parse Text Label and Entity Pairs

procedure PARSE_OUTPUT(output, instruct_word, is_tcree)
    Input: *output* (String), *instruct_word* (String), *is_tcree* (Boolean)
    Output: *text_label* (String), *entity_pairs* (Set of Tuples)

    $instruct\_word \leftarrow$ *instruct_word*
    if $instruct\_word \notin output$ **then**
        return $(\text{""}, \{\})$
    end if
    $text\_label, entity\_pairs \leftarrow$ *output.split(instruct_word, 1)*
    $text\_label \leftarrow text\_label.strip()$
    if *is_tcree* **then**
        $entity\_pairs \leftarrow [entity\_pairs.strip()]$
    else
        $entity\_pairs \leftarrow [pair.strip()$ **for** $pair$ **in** $entity\_pairs.split(\text{" : "})$ **if** $pair]$
    end if
    $entity\_pairs \leftarrow [tuple(pair.split(\text{";"}))$ **for** $pair$ **in** $entity\_pairs]$
    return $(text\_label, \text{set}(entity\_pairs))$
end procedure

Input Output

SCNM: Sentence Classification and Named Entity Recognition Mix Dataset

2018年からクリシューマECで主力メンバーとして活動したが、2020年1月18日、日本のレノファ山口FCへの加入が発表された。/固有表現抽出

文芸/固有表現抽出/:その他の組織名;クリシューマEC:地名;日本:その他の組織名;レノファ山口FC

SCPOS: RW: Sentiment Text Classification and Part-of-Speech: Related Words Mix Dataset

純正の商品なのでとっても使いやすいです。ふたを裏側にしても磁石でつくので邪魔になりません。/感情分析名詞

積極/感情分析名詞/:ポジティブ;純正:ポジティブ;やすい:ポジティブ;邪魔になりません

SCPOS: Adj & N: Sentiment Text Classification and Part-of-Speech: Adjective & Noun Mix Dataset

このバンドを聴いたのは本作が初。印象はというと、オールドロックながらも随所に彼等流の新開拓を施しており、古臭みなく聴ける内容となっている。まだ数回しか聴いてないので断定した言い方はできないが、ハッキリ言ってカッコいい。ロックバンドでも数本の指に入るくらい気に入りましたね。/感情分析形容詞名詞"

積極/感情分析形容詞名詞/:ポジティブ;いい:ポジティブ;カッコいい:中立;たね:ネガティブ;らい:ネガティブ;ドロ:中立;印:中立;回し:中立;開拓:中立;言い方:中立;新:中立;断定:中立;内

SCPOS: Adj: Sentiment Text Classification and Part-of-Speech: Adjective Mix Dataset

個人的にKID Aより好きかもしれない。たぶん一番好き。何よりもわかりやすい。まさに姉妹アルバム。オリジナリティ溢れる作品。かなりロマンティックです。/感情分析形容詞

積極/感情分析形容詞/:ポジティブ;わかりやすい

SCPOS: N: Sentiment Text Classification and Part-of-Speech: Noun Mix Dataset

グリム レジェンド2。映像は綺麗で、ストーリーも良く出来ています。ミニゲームも普通ですね。三作目も作っているそうなので、期待して待っています。/感情分析名詞

積極/感情分析名詞/:中立;ゲーム:ネガティブ;スト:ポジティブ;期待:ポジティブ;出来:ポジティブ;通:中立;普通:中立;目:ポジティブ;良:ポジティブ;綺麗

TCREE: Text Classification and Relation & Event Extraction Mix Dataset

2012年に公開を控える新作映画の中で、MOVIE ENTER編集部オススメの作品を紹介する「2012年先取り映画」。今回“ドラマのサイトー”が紹介する作品は『わが母の記』です。主演の役所広司が出演しています。/関係抽出

映画/関係抽出/:役所広司;主演;わが母の記:

TCONER: Open-domain Text Classification and Named Entity Recognition Mix Dataset

ルネサンス・ソサエティ・オブ・アメリカについて\n\nルネサンス・ソサエティ・オブ・アメリカは1954年に設立され、1300年から1650年の時代の研究を推進しています。RSAは北アメリカや世界中からさまざまな専門分野の学者を集めています。RSAには、大学やカレッジの教授、講師、大学院生、博物館、図書館、他の文化機関のメンバーが5,000人以上おり、独立した学者やルネサンス研究に興味を持つ他の多くの人々もいます。/固有表現抽出

教育/固有表現抽出/:略語;RSA:日付;1954:組織;ルネサンス・ソサエティ・オブ・アメリカ:組織;大学:組織;カレッジ:組織;博物館:組織;図書館:組織;文化機関:職業;教授:職業;講師:職業;大学院生:職業;独立した学者:場所;北アメリカ:場所;世界:主題;ルネサンス研究:数;5,000

Figure 9: The input and output format example with OIELLM in Japanese MRE mix datasets.

Input Output

SCNM: Sentence Classification and Named Entity Recognition Mix Dataset

Since 1989, Sanrio has been using \"Minna no Taabo\" as a character, and in the 1990s, they used Miho Kanno, Mariru Watanabe, Hideyuki Yakou, who appeared in the Hokkaido-based TV drama \"Kita no Kuni Kara,\" and Yoshiji Masuda, who is from Hokkaido, as CM characters./NER

Literature/NER/:Company;Sanrio:Product Name;Minna no Taabo:Person;Miho Kanno:Person;Mariru Watanabe:Location;Hokkaido:Product Name;Kita no Kuni Kara:Person;Hideyuki Yakou:Location;Hokkaido:Person;Yoshiji Masuda

SCPOS: RW: Sentiment Text Classification and Part-of-Speech: Related Words Mix Dataset

A variety of unique numbers are lined up, and it's never boring to listen to. The diversity is wonderful. I think it will remain for future generations./Sentiment related word

positive/Sentiment related word/:neutral;unique:positive;boring:positive;wonderful

SCPOS: Adj & N: Sentiment Text Classification and Part-of-Speech: Adjective & Noun Mix Dataset

Sample dataset in English:\n\ntext: The wolf, who is usually a bad guy, ... The end is cute and heartwarming, and it's a wonderful story. The Japanese version is also wonderful./Sentiment Adj and N

positive/Sentiment Adj and N/:neutral;story:neutral;wolf:positive;wonderful:negative;bad:negative;bad guy

SCPOS: Adj: Sentiment Text Classification and Part-of-Speech: Adjective Mix Dataset

I felt that all the songs had a slow tempo and the melody was hard to grasp. It seems that there were also some songs used as theme songs, but they were not so great and I did not think they were good. I wish there were more understandable melodies. Perhaps, musical preferences vary by individual? I feel like I wasted a little bit of money purchasing it./Sentiment Adj and N

negative/Sentiment Adj and N/:negative;not so great:positive;good:positive;understandable

SCPOS: N: Sentiment Text Classification and Part-of-Speech: Noun Mix Dataset

It contains my favorite songs, and I bought it because it was cheap. It took so long for it to arrive that I thought it would never come, but there is no problem at all with the content. However, it's a minus one because it took so long./Sentiment N

positive/Sentiment N/:neutral;favorite:positive;cheap:negative;problem

TCREE: Text Classification and Relation & Event Extraction Mix Dataset

The top-selling digital camera from October 11th to 16th was Canon's \"IXY 600F\". /relation extraction

IT/relation extraction/:Canon;Product;IXY 600F:

TCONER: Open-domain Text Classification and Named Entity Recognition Mix Dataset

Drama-documentary exploring the betrayals between the Vikings, Anglo-Saxons and Normans. BBC Two/NER

Entertainment/NER/:organization;BBC Two:group;Vikings:group;Anglo-Saxons:group;Normans

Figure 10: The input and output format example with OIELLM in English MRE mix datasets.

Input Output

SCNM: Sentence Classification and Named Entity Recognition Mix Dataset

几乎在同一时间，村田真等人也在开发与XML Schema不同的新的模式语言RELAX。/实体命名识别

技术/实体命名识别/:人名;村田真:产品名;XML Schema:产品名;RELAX

SCPOS: RW: Sentiment Text Classification and Part-of-Speech: Related Words Mix Dataset

科林·费尔斯赢得了最佳男主角奖，然而此次获奖是因为克里斯蒂安·贝尔是配角，如果他是主角，那么这部电影可能就是毫无存在感的，这部电影根本不值得一提，是部烂片。/感情分析关联单词

消极/感情分析关联单词/:消极;烂片"

SCPOS: Adj & N: Sentiment Text Classification and Part-of-Speech: Adjective & Noun Mix Dataset

虽然不用说内容，但能在美丽的画面和声音中观看，感到很开心。/感情分析形容词名词

积极/感情分析形容词名词/:积极;开心:中立;观看:中立;内容:积极;美丽

SCPOS: Adj: Sentiment Text Classification and Part-of-Speech: Adjective Mix Dataset

虽然是机械式的，但时间被准确地标记，分钟的显示在外面，所以可以快速确认分钟，设计简单轻巧实惠，很好。/感情分析形容词名词

积极/感情分析形容词名词/:积极;好

SCPOS: N: Sentiment Text Classification and Part-of-Speech: Noun Mix Dataset

对于朱莉亚罗伯茨的粉丝来说，这部电影简直让人无法抗拒，休·格兰特也很棒，无论看多少遍都很享受。/感情分析名词

积极/感情分析名词/:积极;棒:积极;粉丝:积极;享受

TCREE: Text Classification and Relation & Event Extraction Mix Dataset

11日深夜，日本电视台的“Going! Sports＆News”节目中，上次参加泳泳太平洋锦标赛时展现出绝对实力，在100米蛙泳和200米蛙泳项目中获得冠军的北岛康介进行了现场直播。/事件抽取

运动/事件抽取/:北岛康介;出演;日本电视台;11日:

TCONER: Open-domain Text Classification and Named Entity Recognition Mix Dataset

预制结构，如所谓的\"移动\"住宅或模块化住宅等，通常通过放置在混凝土或砖块基座上进行安装。通常，这些结构的外围下沿是不受支撑的。这些边缘与地面之间的间隙可能被金属或合成材料制成的非承重裙子覆盖，仅仅是为了提供更加美观的外观。经过很长一段时间，这些建筑的外围边缘可能会翘曲、下垂或变形。裙子对阻止这种形式的恶化并不具备结构上的足够性。/实体命名识别

建筑/实体命名识别/:位置;地面:状态;恶化:结构;模块化住宅:结构;裙子:结构;外围边缘:材料;混凝土:材料;砖块:材料;金属:材料;合成材料

Figure 11: The input and output format example with OIELLM in Chinese MRE mix datasets.