A Multilingual Dataset and Empirical Validation for the Mutual Reinforcement Effect in Information Extraction

Anonymous ACL submission

Abstract

The Mutual Reinforcement Effect (MRE) describes a phenomenon in information extraction where word-level and sentence-level tasks can mutually improve each other when jointly modeled. While prior work has reported MRE in Japanese, its generality across languages and task settings has not been empirically validated, largely due to the lack of multilingual MRE datasets. To address this limitation, we introduce the Multilingual MRE Mix dataset (MMM), which consists of 21 sub-datasets covering English, Japanese, and Chinese. We propose an LLM-assisted dataset translation and alignment framework that significantly reduces manual annotation effort while preserving the structural requirements of MRE tasks. Building on MMM, we adopt a unified input-output framework to train an open-domain information extraction model and conduct extensive empirical studies, including full fine-tuning ablations and the construction of knowledgeable verbalizers based on MRE-mix data. Experimental results show that 76 percent of the MMM sub-datasets consistently exhibit the Mutual Reinforcement Effect across languages. These findings provide systematic empirical validation of MRE in multilingual settings and demonstrate its practical value for information extraction.

1 Introduction

Information extraction (IE) Sarawagi et al. (2008) is a fundamental research area in natural language processing (NLP), aiming to transform unstructured text into structured representations. Over the years, IE has evolved into a collection of specialized 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),

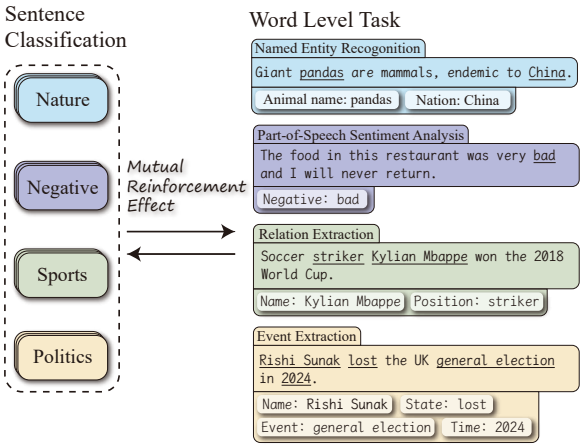

Figure 1: The Mutual Reinforcement Effect between word-level labels and a text-level label within the same text. **A word-level IE task is regarded as a Point, and a text-level IE task is regarded as a Line. Mutual Reinforcement Effect exists between the Point and the Line.**

relation 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 subtasks have been studied and modeled in isolation.

Multi-task learning for IE (Sun et al., 2023; Zhao et al., 2020) attempts to bridge this separation by jointly training multiple subtasks within a single model. In most existing approaches, datasets from different IE tasks are merged, and the model is optimized with task-specific output heads. While this paradigm allows models to share representations across tasks, it treats each task as an independent objective and does not explicitly model or analyze the semantic interactions between tasks. As a result, whether and how different IE subtasks can mutually benefit from each other remains underexplored.

The Mutual Reinforcement Effect (MRE) (Gan et al., 2023b) was proposed to explicitly study this interaction. MRE refers to the phenomenon that

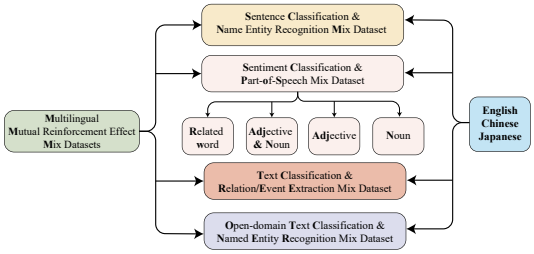

Figure 2: Overview of the Multilingual Mutual Reinforcement Effect Mix (MMM) datasets and their sub-datasets.

word-level IE tasks and text-level IE tasks can mutually enhance each other when they are jointly modeled on the same text. In this formulation, text-level tasks such as sentence classification or sentiment analysis provide global semantic context, while word-level tasks such as NER provide fine-grained semantic signals. Rather than treating them as parallel objectives, MRE emphasizes their bidirectional dependency within a single input instance.

To illustrate this idea, IE tasks are categorized into word-level tasks and text-level tasks, which are abstracted as Points and Lines, respectively. Understanding the Point facilitates understanding the Line, and vice versa. Figure 1 presents a concrete example of this interaction. The sentence "Giant pandas are mammals, endemic to China." is labeled as *nature* at the text level, while containing word-level entity annotations such as *Animal Name: pandas* and *Nation: China*. The sentence-level label constrains the plausible entity types, while the recognized entities reinforce the semantic correctness of the sentence-level classification. This interaction is not task-specific but reflects a general mechanism of human language understanding, where global meaning is inferred from words and words are interpreted under global context (Gan et al., 2023c).

Although MRE has been shown to be effective in prior studies, its empirical exploration has been severely constrained by data availability. Existing MRE mix datasets are limited to Japanese, which restricts systematic investigation across languages and hinders broader validation of the effect. As a consequence, it remains unclear whether MRE is language-specific or a more general property of IE tasks.

To address this limitation, we construct the Multilingual Mutual Reinforcement Effect Mix (MMM) datasets, covering English, Chinese, and Japanese. As shown in Figure 2, MMM consists of 21 sub-datasets, with seven datasets per language. These datasets span multiple MRE task combinations, including sentence classification with NER (SCNM), sentiment classification with part-of-speech related supervision (SCPOS), text classification with relation and event extraction (TCREE), and an open-domain text classification and NER dataset (TCONER). To enable this expansion, we propose an LLM-assisted dataset translation framework that substantially reduces manual annotation cost while preserving annotation consistency.

Based on MMM, we conduct a series of empirical studies to systematically validate the MRE hypothesis. First, we design a unified input-output format and train an Open-domain Information Extraction Large Language Model (OIELLM) on MMM, demonstrating that models trained with MRE mix data achieve stronger and more stable IE performance. Second, we perform controlled ablation experiments across all 21 sub-datasets. The results show that 76% of the datasets exhibit clear positive reinforcement between word-level and text-level tasks, providing direct empirical evidence of MRE across languages and task types. Finally, we incorporate word-level supervision from MRE datasets into a Knowledgeable Verbalizer framework (Hu et al., 2022) for text-level classification. The observed performance gains further support the claim that word-level information encoded by MRE contributes meaningfully to text-level inference.

In summary, this work focuses on empirically validating the Mutual Reinforcement Effect rather than proposing a new model architecture. Our main contributions are as follows:

1. We construct the first multilingual MRE mix dataset, MMM, extending existing Japanese-only resources to English and Chinese and covering 21 sub-datasets across multiple IE task combinations.

2. We provide a systematic empirical validation of the Mutual Reinforcement Effect through large-scale ablation studies across languages, showing that MRE is a robust and reproducible phenomenon rather than a language-specific artifact.

3. We demonstrate the practical utility of MRE by applying word-level supervision to a

Knowledgeable Verbalizer framework, offering additional evidence that MRE enhances text-level IE tasks beyond joint training.

2 Related Work

Datasets. Existing MRE mix datasets originate from Japanese-only resources, including SCNM Gan et al. (2023b), SCPOS Gan et al. (2023d), and TCREE Gan et al. (2023a). While these datasets have demonstrated the feasibility of Mutual Reinforcement Effect, their exclusive focus on Japanese has limited broader empirical validation and cross-lingual exploration of MRE.

In parallel, recent studies have increasingly leveraged Large Language Models (LLMs) for dataset construction and annotation (Tan et al., 2024; Wadhwa et al., 2023; Li et al., 2023; Laskar et al., 2023). Prior work has shown that LLM-assisted annotation can substantially reduce human effort while maintaining competitive quality (Huang et al., 2023). Such approaches have been applied to diverse domains, including mathematical reasoning datasets (Lin et al., 2024) and iterative annotation frameworks such as FreeAL (Xiao et al., 2023a). These methods typically rely on instruction learning and in-context learning to guide LLMs to produce structured labels from unstructured inputs.

Different from prior work, the MMM dataset is designed specifically to support empirical investigation of MRE across languages. By translating existing MRE mix datasets into English and Chinese and expanding them with an open-domain NER dataset (TCONER), MMM enables systematic cross-lingual validation of word-level and text-level task interactions under a unified framework.

LLMs for Information Extraction. Generative IE models based on sequence-to-sequence architectures have become increasingly prominent, starting from UIE Lu et al. (2022) and extending to models such as USM Lou et al. (2023) and Mirror (Zhu et al., 2023). These approaches unify multiple word-level IE tasks, including NER, relation extraction, and event extraction, through standardized input-output formats, allowing a single model to handle diverse IE subtasks.

With the emergence of LLMs, IE research has largely followed two directions. One line of work directly queries LLMs in zero-shot or few-shot settings using carefully designed prompts (Wang et al., 2023; Wei et al., 2023). Another line focuses on fine-tuning LLMs with task-specific or instruction-based datasets to improve extraction accuracy (Zhou et al., 2023; Xiao et al., 2023b). While these studies demonstrate strong IE capabilities, they primarily treat IE subtasks as independent objectives.

In contrast, our work centers on the Mutual Reinforcement Effect, which explicitly studies the bidirectional interaction between word-level and text-level IE tasks within the same input. Rather than proposing a new generative IE architecture, we use MMM as an empirical testbed to examine whether and how such interactions consistently emerge across languages and task combinations.

3 Multilingual Mutual Reinforcement Effect Mix Datasets

This section introduces the construction of the Multilingual Mutual Reinforcement Effect Mix (MMM) datasets. Our goal is not to fully automate dataset creation, but to design a practical human–LLM collaborative framework that reduces repetitive manual translation while preserving annotation quality through systematic rule-based filtering and human verification.

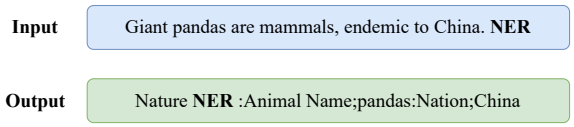

Figure 3: The format of MMM datasets.

3.1 Dataset Translation Framework

We first briefly introduce the MMM dataset format. As shown in Figure 3, each sample consists of an input and an output. The input includes the raw text together with a task instruction (e.g., NER), while the output contains a text-level label followed by word-level label–entity pairs. The output format follows the original MRE design, using structured delimiters (e.g., ":", ";") to ensure consistency across different IE subtasks and languages.

Figure 4 illustrates the overall translation workflow. We begin by applying rule-based matching to the original Japanese MRE datasets. Since label sets are fixed and shared across datasets, all text-level and word-level labels are translated deterministically using predefined mappings (e.g., "ポジティブ" → "positive"). This step is fast, precise, and removes ambiguity before involving LLMs.

The translated labels are then combined with the original text and entity spans and provided to an

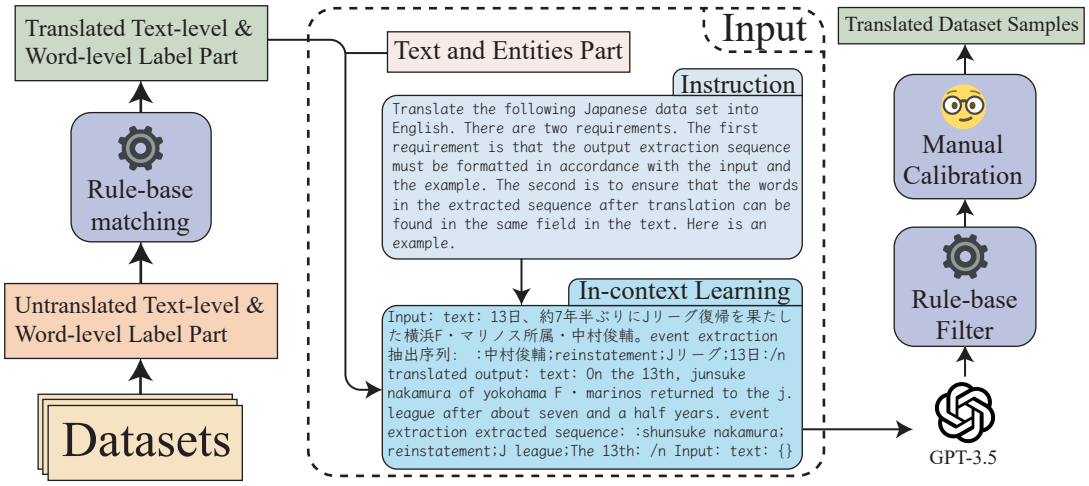

Figure 4: The overview of the dataset translation framework.

LLM, GPT-3.5-Turbo (Ouyang et al., 2022), which is used solely to assist in translating the remaining free-text content. We adopt instruction-based prompting together with one-shot in-context learning to guide the model. Instead of optimizing for fully automatic translation accuracy, the prompts explicitly constrain the output format and emphasize entity span consistency with the translated text.

Due to the inherent variability of LLM outputs, we apply a two-stage rule-based filtering process. First, samples containing untranslated Japanese characters are removed. Second, samples with entity spans that cannot be aligned with the translated text or violate MMM formatting constraints are discarded. This filtering step intentionally favors precision over coverage.

The remaining samples are then manually reviewed and calibrated by ten multilingual graduate students proficient in at least two of Chinese, English, and Japanese. Human annotators focus on correcting terminology, resolving minor inconsistencies, and validating domain-specific expressions using external references such as dictionaries. In practice, this framework significantly reduces repetitive translation work, while reserving human effort for quality control rather than raw translation.

Although GPT-3.5-mini serves as the backbone of the pipeline, its role is strictly supportive. Empirically, more advanced models such as GPT-4o did not consistently reduce manual correction effort and were therefore not cost-effective in this setting. Comparative analyses are reported in Appendix E. Overall, the proposed framework is best viewed as a scalable annotation aid rather than a fully automated translation system.

4 Open-domain Information Extraction Large Language Model

This section introduces the Open-domain Information Extraction Large Language Model (OIELLM), which is specifically designed to support empirical investigation of the Mutual Reinforcement Effect on MMM datasets. Rather than proposing a new architecture, OIELLM focuses on a unified input-output formulation that enables large language models to jointly generate text-level labels and word-level label–entity pairs within a single decoding process.

MRE mix datasets differ fundamentally from conventional IE benchmarks. Each sample requires the model to output both a global text-level label and multiple word-level extractions from the same input. Standard sequence labeling models cannot directly support this requirement, while most existing generative IE frameworks primarily target word-level outputs such as entities, relations, or events, without explicitly modeling text-level supervision.

The core objective of OIELLM is to operationalize MRE by enforcing a shared generation space for both levels of supervision. By learning to predict text-level labels and word-level label–entity pairs jointly, the model is encouraged to capture their bidirectional dependency, which forms the basis of MRE. This formulation also improves interpretability, as the generated word-level evidence directly supports the predicted text-level label.

Instead of adopting dialog-style question-answer prompting, we follow prior generative IE paradigms and design a task-agnostic yet structured input-output format tailored to MMM datasets. Fig-

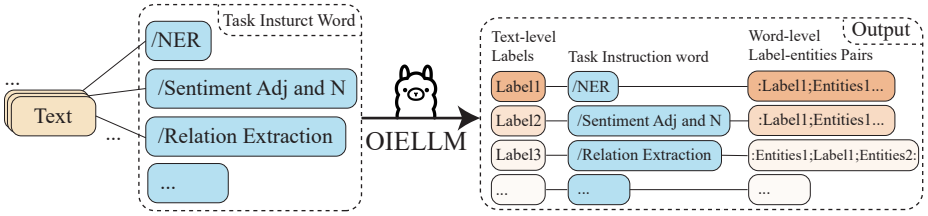

Figure 5: The input and output of Open-domain Information Extraction Large Language Model (OIELLM).

ure 5 illustrates the overall design. Each input consists of the raw text augmented with one or more task instruction words. These instruction words are explicitly marked with a special prefix symbol "/" to distinguish them from the text content.

The output follows a fixed and unified structure. It begins with text-level labels, followed by the corresponding word-level label–entity pairs associated with each task instruction. We retain the delimiter-based format (":", ";") from prior MRE datasets to ensure consistency and unambiguous parsing across tasks and languages.

This format design is critical. By removing dialog prompts and standardizing task specification through instruction tokens, OIELLM reduces input length, avoids prompt-induced variance, and allows the model to focus on learning structural dependencies between text-level and word-level information. Importantly, the same format is shared across all MMM sub-datasets, enabling a single model to process diverse IE tasks under a unified generation framework.

In summary, OIELLM is not merely a fine-tuned LLM, but a structured generation framework that makes MRE observable and measurable across multiple IE subtasks and languages.

5 Experiment

This section describes the experimental setup used to empirically evaluate the Mutual Reinforcement Effect, including model training details and evaluation protocols.

5.1 Details of OIELLM Training

We select USA-7B (Instruction + In-context Learning)[1] and GIELLM-13B-jp[2] as baselines, as they are the only prior models explicitly trained on MRE mix datasets. For OIELLM, we adopt Meta-LLaMA3-8B-Instruct[3] as the primary backbone. Since LLaMA3 does not provide a 13B variant, we additionally include LLaMA2-13B Touvron et al. (2023) for comparison.

We also tested GPT-3.5-Turbo and GPT-4o-mini under one-shot instruction and in-context settings. However, these models failed to consistently generate outputs that conform to the required structured format of MRE mix datasets. Even with sufficient demonstrations, their outputs frequently violated formatting constraints or merged multiple fields, resulting in near-zero F1 scores. As these failures reflect format adherence rather than semantic understanding, we do not include these models as competitive baselines in the main experiments.

OIELLM is fully fine-tuned with all parameters updated. Training is conducted in BF16 precision and inference in FP16. Models are trained for three epochs with a learning rate of 1×10^{-5}, using three A800 80GB GPUs and three RTX 6000 Ada 48GB GPUs. Training time ranges from 12 to 20 hours depending on model size. Dataset statistics for training and evaluation splits are reported in Appendix Tables 5 and 6.

5.2 Evaluation

We adopt F1 score as the primary evaluation metric, following a strict structured prediction protocol. Model outputs are first separated into text-level labels and word-level label–entity pairs according to task instruction words. Word-level outputs are parsed using predefined delimiters (":", ";") and evaluated as unordered sets.

We report three metrics: Text-Level F1 (TL), Word-Level F1 (WL), and ALL, where ALL measures correctness only when both text-level and word-level outputs are simultaneously correct within a single prediction.

This strict evaluation is intentional. MRE mix datasets are designed as structured information extraction tasks, where outputs must be directly us-

[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

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	81.20	2.59	1.40	1.33	3.01	0
GPT-4o	58.30	23.42	8.57	87.00	4.35	1.30	82.33	0.49	0.03
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
GPT-4o	68.90	0.21	0.17	74.17	0.36	0.03	90.43	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
GPT-4o	44.63	0	0	86.58	3.97	0.99	85.00	17.43	0.47
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
GPT-4o	24.00	1.67	1.07	16.77	0.92	0.03	13.90	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
GPT-4o	67.43	26.11	13.47	60.38	10.91	0.99	79.27	3.69	0.50
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
GPT-4o	41.60	2.00	0.83	83.93	1.45	0.47	41.60	2.00	0.83
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
GPT-4o	30.40	5.77	0.03	36.30	9.13	0.03	40.50	9.02	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:

able as machine-readable annotations. Even minor deviations in formatting, missing fields, or merged entities invalidate the extraction. While general-purpose LLMs often produce semantically reasonable text, they frequently fail to meet these structural constraints, which explains their low scores

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.

under this protocol. Detailed evaluation procedures and formulas are provided in Appendix G.

6 Results

Table 1 reports the performance of three OIELLM variants trained on the 21 MMM sub-datasets. The model marked with an asterisk, OIELLM-8B, is initialized from LLaMA3-8B-Instruct, while the remaining models use the LLaMA3-8B-Base backbone. Overall, OIELLM achieves strong and stable performance across languages and tasks.

Notably, OIELLM outperforms GIELLM-13B-jp on approximately half of the Japanese datasets, despite GIELLM being specifically designed for Japanese. This result suggests that combining multilingual supervision with MRE-style multitasking can more effectively activate and reuse knowledge embedded in multilingual pretrained LLMs, even for a single target language.

Performance on the TCONER datasets is relatively weaker. We attribute this primarily to data scarcity, as open-domain tasks require substantially larger and more diverse training data than domain-specific settings. Addressing this limitation will be part of future work.

We additionally evaluated GPT-3.5-Turbo and GPT-4o-mini under one-shot instruction and in-context learning settings. Their F1 scores are consistently low under our evaluation protocol. This is mainly due to two factors: (1) the strict structured-output evaluation, where even minor formatting deviations invalidate predictions, and (2) the absence of supervised fine-tuning for MRE-style joint text-level and word-level extraction. These results further motivate the need for dedicated IE models trained explicitly on MRE mix datasets.

7 Ablation Experiment of MMM Datasets

Details of the ablation setup are provided in Appendix A and Appendix B. Table 2 summarizes the results.

Across the six fixed-label datasets, models trained with additional level information consistently outperform their counterparts trained without it. Overall, **76% of the ablation results show that incorporating information from one level facilitates performance at the other level**. This provides strong empirical evidence for the Mutual Reinforcement Effect, confirming that word-level and text-level supervision mutually enhance each other when jointly modeled.

These findings support the central hypothesis of this paper: balanced integration of text-level and word-level tasks improves model understanding and extraction performance, reflecting a reinforcement mechanism analogous to human language comprehension, where local lexical cues and global semantic judgments inform each other.

For open-domain text classification and NER, the results are more mixed. Some datasets contain multiple or weakly correlated labels, in which not all word-level information contributes positively to text-level prediction. This reduces the overall reinforcement effect. However, for the Chinese and Japanese TCONER datasets, incorporating level information consistently improves performance. This observation suggests that MRE may be more effective in character-based languages, where lexical units often carry richer semantic information, compared to alphabetic languages such as English.

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

These ablation results establish a clear empirical foundation for further analysis of how word-level information can be selectively leveraged to enhance text-level inference, which we explore next through Knowledgeable Verbalizer experiments.

8 Word-level Information as Knowledgeable Verbalizer

To examine whether the Mutual Reinforcement Effect (MRE) extends beyond information extraction, we conduct an additional empirical study on few-shot text classification. In the MRE framework, word-level information (WLI) is expected to support text-level understanding. Accordingly, we use WLI extracted from MRE mix datasets to construct Knowledgeable Verbalizers (KV) (Hu et al., 2022).

For each classification label, we select the top-100 high-frequency words from the corresponding word-level annotations as label-specific verbalizers. During inference, token probabilities are aggregated at the label level, and the label with the highest aggregated probability is predicted. Compared with conventional KV construction based on general-purpose related-word resources, WLI-based KVs are task-aligned and dataset-specific. Superior performance of WLI-based KVs thus provides indirect evidence that word-level information facilitates text-level classification, supporting the MRE hypothesis beyond IE tasks. Experimental details and examples are provided in Appendix B and Appendix Figure 8.

8.1 Results of Word-level Information as Knowledgeable Verbalizer

Table 3 shows KV-based text classification results on 18 MMM sub-datasets in English, Chinese, and Japanese. The open-domain TCONER dataset is excluded due to its unfixed label schema.

WLI-based KVs achieve the best performance on 16 out of 18 datasets and consistently outperform the original KV construction method. Gains are especially notable on sentiment classification tasks, where word-level polarity is crucial. Unlike the ablation experiments that evaluate MRE within IE tasks, these results demonstrate that WLI learned through MRE transfers effectively to downstream text classification, providing additional empirical evidence for the existence and practical utility of MRE.

9 Conclusion and Future Work

This paper presents a large-scale empirical study of the MRE in information extraction. We construct the MMM datasets by translating existing Japanese MRE datasets into English and Chinese and by introducing the TCONER dataset to cover open-domain IE tasks. To reduce annotation cost, we adopt an LLM-assisted translation framework that supports, rather than replaces, human annotation.

Based on the MMM datasets, we train an OIELLM with a unified input-output format that jointly models text-level and word-level information. Extensive experiments show that OIELLM achieves strong multilingual IE performance. More importantly, ablation studies on 21 sub-datasets demonstrate that in 76% of cases, information at one level facilitates learning at the other level, providing direct empirical evidence for the MRE hypothesis.

We further apply word-level information as Knowledgeable Verbalizers in few-shot text classification and observe consistent improvements over conventional KV construction methods. These results indicate that MRE-derived word-level information can effectively enhance downstream text-level tasks, forming a threefold empirical validation of MRE across ablation analysis, IE modeling, and downstream application.

Future work will extend MRE to more languages and task combinations and explore incorporating MRE-aware objectives into pre-training and instruction tuning of large language models.

10 Limitations

This study has several limitations worth noting. First, although we experimented with different large language models for dataset translation, our empirical results show that GPT-4o does not consistently outperform GPT-3.5-mini in the context of MRE dataset translation. Due to budget constraints, we did not exhaustively evaluate all available models across all sub-datasets.

Second, the proposed dataset translation framework is not designed to fully automate dataset construction. Instead, it functions as an annotation-assistance tool that reduces repetitive translation effort for relatively simple cases, while leaving complex or ambiguous instances to human annotators. All translated samples are therefore manually verified and refined. As a result, we do not report automatic translation quality metrics, and instead provide the proportion of manually corrected samples as a coarse indicator of translation reliability.

Finally, while our empirical results demonstrate the Mutual Reinforcement Effect across multiple tasks and languages, the current study focuses on specific combinations of text-level and word-level information. Extending MRE to other task configurations and exploring automatic criteria for identifying beneficial task interactions remain open directions for future work.

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.

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].

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

B 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 training set was utilized to fully parameterize the fine-tuned LLMs. Subsequently, 1,000 samples were

[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

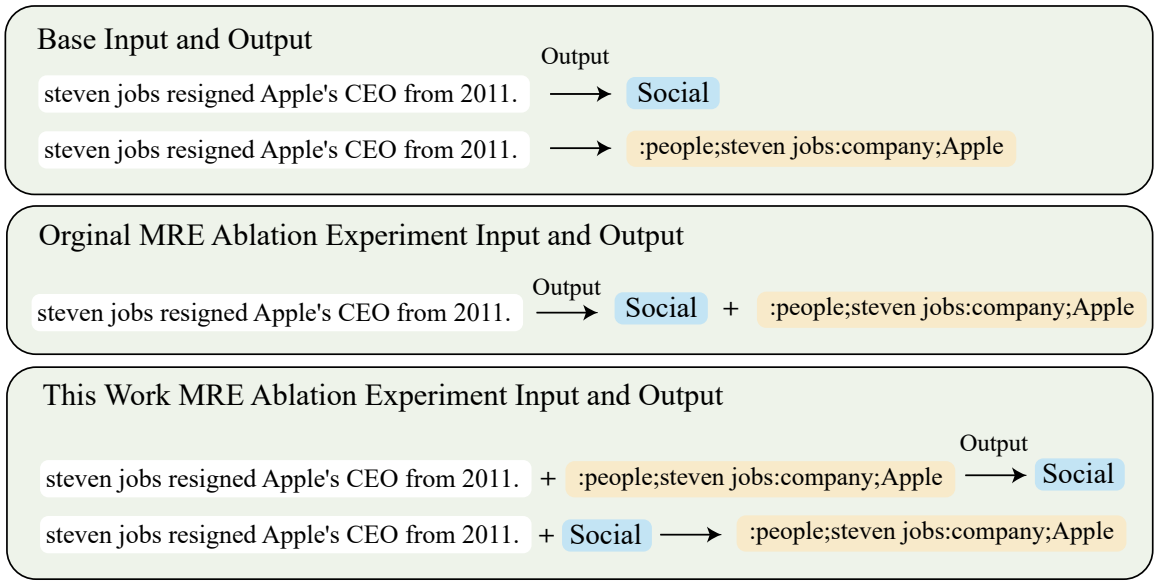

Figure 6: The figure shows the inputs and outputs of the traditional ablation experiment for the MRE task and the new empirical MRE experiment proposed in this work.

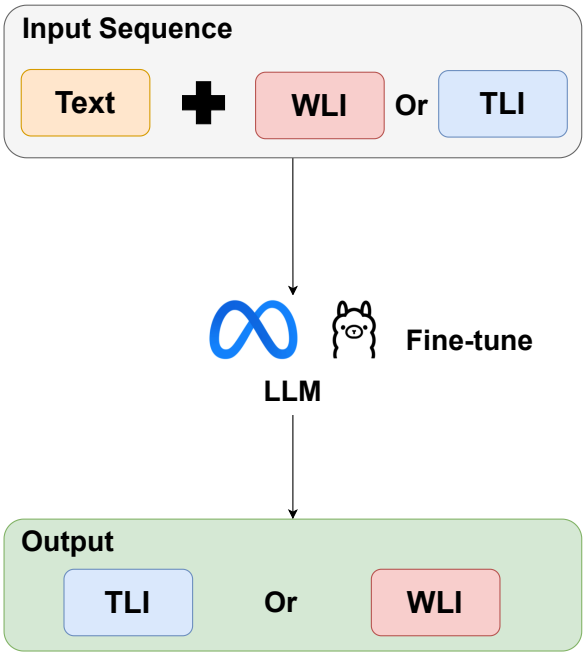

Figure 7: The figure illustrates the flow of an empirical MRE experiment using the new approach.

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.

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 4: 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.)

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

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

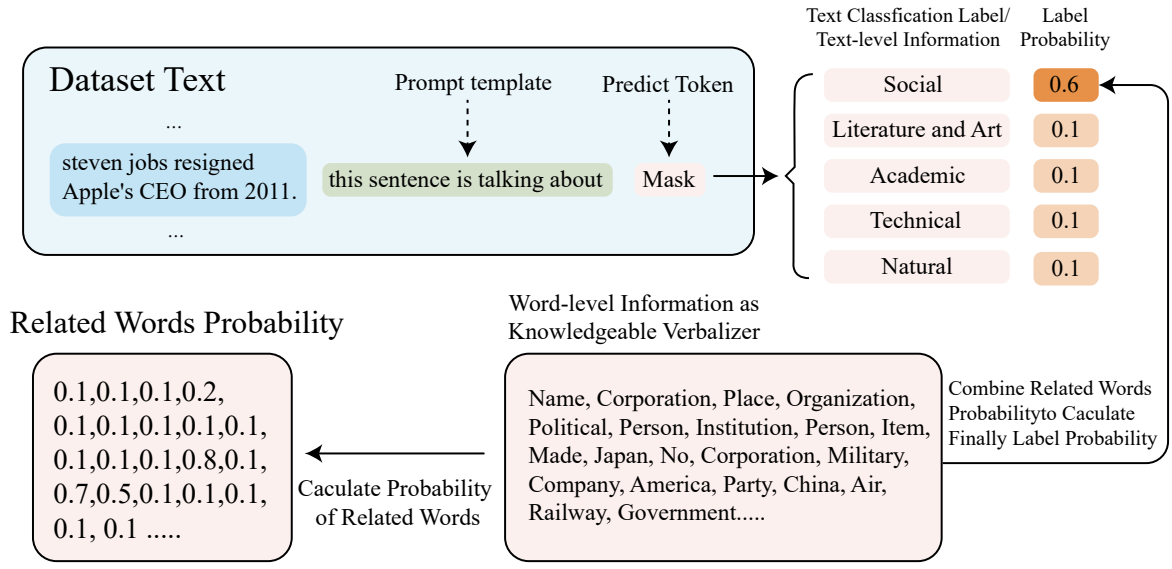

Figure 8: The figure demonstrates how word-level information is utilized as a Knowledgeable Verbalizer to assist in text-level classification tasks. Additionally, it provides a detailed explanation of the functioning of the Knowledgeable Verbalizer.

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.

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 4 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 5 and 6, the statistics for the complete training and test sets of the MMM dataset.

[10] https://chatgpt.com/

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

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 5: Statistical results of train sets of OIELLM.

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 6: Statistical results of test sets.

The MMM dataset was segmented into 21 sub-datasets. Training set sizes were assigned based on the sizes of these sub-datasets, categorized into three groups: 500, 1000, and 2000 samples. Samples beyond these numbers were allocated to the test sets.

E Comparison of GPT-4o and GPT-3.5-mini on Dataset Translation Frameworks

In our translation pipeline, we initially selected GPT-3.5-Turbo (referred to as GPT-3.5-mini) for translating Japanese datasets due to its favorable trade-off between translation quality and cost-efficiency. Although GPT-4o is a more advanced model, our empirical tests suggest that it did not consistently outperform GPT-3.5-mini in our use case, particularly for domain-specific and structured datasets relevant to Mutual Reinforcement Effect (MRE) tasks.

To validate this decision, we conducted a comparative evaluation of both models across multiple datasets. The workflow included two rounds of rule-based filtering to remove untranslated Japanese text and span mismatches, followed by two rounds of manual review. We report the proportion of samples retained after rule-based filtering, which serves as an indirect indicator of translation consistency and quality. The results are summarized in Table 7.

As shown, GPT-4o exhibited slight improvements in a few datasets (e.g., TCREE) but also underperformed in others (e.g., SCNM and TCONER), resulting in a higher sample loss percentage after filtering. These results support our initial choice of GPT-3.5-Turbo as a more cost-effective and stable translation backbone.

Despite this, we recognize that future improvements in translation models could further reduce manual correction efforts. For instance, although our current framework already reduced 20–40% of samples through rule-based filtering and required manual correction in only 10% of the remaining data, employing more accurate models may streamline this process even further. We plan to incorporate GPT-4o or its successors in subsequent updates of the MMM dataset.

F UI of the Tools for Manually Correcting Translated Dataset

As illustrated in the figure 12, the top section of the UI interface allows users to select two dataset files: the original Japanese dataset and its translated counterpart in either English or Chinese. Once selected, corresponding samples from both the original and translated datasets are displayed below for direct comparison. Users can manually edit the translations as needed and apply corrections by clicking the "Fix" button. If the translation is accurate, the user can proceed to the next sample by clicking "Next," streamlining the verification process.

G Calculate Detail of F1 Score

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

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

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

Dataset	Direction	Model	Remaining Samples	Lost %
SCNM	JA → EN	GPT-3.5-Turbo	4449	16.73%
SCNM	JA → EN	GPT-4o	3996	25.21%
SCPOS: RW	JA → EN	GPT-3.5-Turbo	1312	34.40%
SCPOS: RW	JA → EN	GPT-4o	1182	41.04%
SCPOS: Adj&N	JA → EN	GPT-3.5-Turbo	4801	51.99%
SCPOS: Adj&N	JA → EN	GPT-4o	4821	51.79%
TCREE	JA → EN	GPT-3.5-Turbo	1910	4.50%
TCREE	JA → EN	GPT-4o	1979	1.05%
TCONER	EN → ZH	GPT-3.5-Turbo	9047	9.53%
TCONER	EN → ZH	GPT-4o	7951	20.49%

Table 7: Translation Quality Comparison between GPT-3.5-Turbo and GPT-4o

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 8: 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)

H 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.

Figure 12: UI of the Tools for Manually Correcting Translated Dataset.