# Building a Japanese Document-Level Relation Extraction Dataset Assisted by Cross-Lingual Transfer

**Youmi Ma, An Wang, Naoaki Okazaki**
Tokyo Institute of Technology
Tokyo, Japan
{youmi.ma@nlp., an.wang@nlp., okazaki@}c.titech.ac.jp

**Abstract**

Document-level Relation Extraction (DocRE) is the task of extracting all semantic relationships from a document. While studies have been conducted on English DocRE, limited attention has been given to DocRE in non-English languages. This work delves into effectively utilizing existing English resources to promote DocRE studies in non-English languages, with Japanese as the representative case. As an initial attempt, we construct a dataset by transferring an English dataset to Japanese. However, models trained on such a dataset suffer from low recalls. We investigate the error cases and attribute the failure to different surface structures and semantics of documents translated from English and those written by native speakers. We thus switch to explore if the transferred dataset can assist human annotation on Japanese documents. In our proposal, annotators edit relation predictions from a model trained on the transferred dataset. Quantitative analysis shows that relation recommendations suggested by the model help reduce approximately 50% of the human edit steps compared with the previous approach. Experiments quantify the performance of existing DocRE models on our collected dataset, portraying the challenges of Japanese and cross-lingual DocRE.

**Keywords:** Information Extraction, Document-level Relation Extraction, Dataset Construction, Japanese

## 1. Introduction

Document-level Relation Extraction (DocRE) aims to identify all semantic relationships between entities in a document (Yao et al., 2019). The task promotes Relation Extraction (RE) to a more practical setting, where relations can reside between entity pairs *document-wise*, i.e., within and beyond the sentence boundary. DocRE is worth spotlighting as it not only inherits the significance of RE in benefiting knowledge graph completion and question answering but also showcases how models comprehend long text (Yu et al., 2017; Trisedya et al., 2019; Chen et al., 2023a). Even in the era of large language models (LLMs), the task deserves more attention as in-context learning of DocRE was considered not yet feasible (Wadhwa et al., 2023).

DocRE research has been conducted mainly in English (Yao et al., 2019; Zhou et al., 2021; Tan et al., 2022b). This work aims to promote DocRE in other languages with the help of English resources. Specifically, we utilize existing resources of English DocRE to construct datasets and models for non-English DocRE. We chose Japanese as our target language for the following two reasons. Firstly, despite Japanese being a widely used language for web content, there is currently a notable absence of general-purpose Japanese DocRE resources. Our work thus contributes to the community by establishing the foundation for Japanese DocRE. Secondly, Japanese stands out as one of the most linguistically distant languages from English (Chiswick and Miller, 2004). The dissimilarity encompasses various aspects, including script models and word order. Therefore, our research setting is highly representative, and the insights we gain will hold value when acquiring resources for other languages.

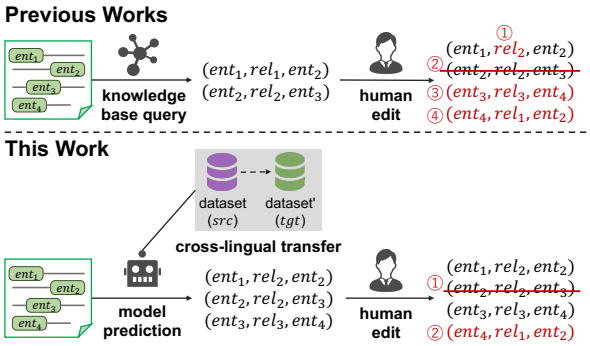


Figure 1: Overview of the proposed annotation scheme. *src* and *tgt* represent the source and target language, respectively. Previous works require 4 human edit steps to reach the final annotation, while ours only require 2.

We first explore if DocRE resources of high quality can be obtained with zero human effort. To this end, we automatically construct a Japanese DocRE dataset with cross-lingual transfer. Specifically, we translate Re-DocRED (Tan et al., 2022b), a popular English DocRE dataset of high quality, into Japanese with a machine translator. An automatically constructed dataset (hereafter referred



*LREC-COLING 2024*, pages 2567–2579

| Dataset | Lang. | # Triples | # Docs. | Avg. # Toks. | # Rels. | Evi. |
|---|---|---|---|---|---|---|
| DocRED (Yao et al., 2019) | *en.* | 50,503 | 4,051 | 198.4 | 96 | Y |
| Re-DocRED (Tan et al., 2022b) | *en.* | 120,664 | 4,053 | 198.4 | 96 | N |
| HacRED (Cheng et al., 2021) | *zh.* | 56,798 | 7,731 | 122.6 | 26 | N |
| HistRED (Yang et al., 2023) | *kr.* | 9,965 | 5,816 | 100.6 | 20 | Y |
| **JacRED** (Ours) | *ja.* | 42,241 | 2,000 | 260.1 | 35 | Y |

Table 1: Statistics of existing and proposed DocRE datasets. Column **Evi.** shows whether each dataset annotates evidence sentences or not. Statistics for DocRED are from the human-annotated subset.

to as *Re-DocRED*$^{ja}$) can thus be obtained without human annotators. The translation-based cross-lingual transfer has been successfully applied to other information extraction (IE) tasks, including named entity recognition and sentence-level relation extraction (Chen et al., 2023b; Hennig et al., 2023). However, we observe that models trained on Re-DocRED$^{ja}$ suffer from low recalls when extracting relation triples from raw Japanese text. We investigate the error cases and attribute the failures to the discrepancies between documents in Re-DocRED$^{ja}$ and those composed by native speakers. The discrepancies include deviations of topics and wording. Our observation underscores the uniqueness and complexity of DocRE in comparison to other IE tasks.

Given that Re-DocRED$^{ja}$ is not suitable for immediate practical application, we explore if the dataset can assist human annotation. As in Figure 1, we adopt a semi-automatic, edit-based annotation scheme, where annotators edit machine recommendations by removing incorrect instances and supplementing missed instances (Yao et al., 2019; Cheng et al., 2021; Tan et al., 2022b). In contrast to previous works where only relation instances from an existing knowledge base are recommended (Yao et al., 2019; Cheng et al., 2021), we recommend instances with a state-of-the-art DocRE model trained on Re-DocRED$^{ja}$. The collected dataset is named as **JacRED** (Japanese Document-level Relation Extraction Dataset), with statistics shown in Table 1. We quantitatively analyze recommendations from the model trained on Re-DocRED$^{ja}$ and those from knowledge base queries and find the former reduces the human edit steps to half of the latter.

We employ JacRED as a benchmark for evaluation. Firstly, we evaluate the performance of existing models on Japanese DocRE. While models trained using the train set of JacRED perform fairly on the test set, the scores fall short of those achieved on Re-DocRED. The result indicates that JacRED introduces extra challenges in addition to Re-DocRED. Notably, we observe that in-context learning of LLMs yields poor performance on JacRED, in line with the findings of Wadhwa et al. (2023). Next, we quantify the performance gap between models trained on Re-DocRED$^{ja}$ and those trained on JacRED. The results further demonstrate that, although translation-based cross-lingual transfer appears effective for a range of IE tasks, it does not hold true for DocRE, especially for distant language pairs. Additionally, JacRED also enables the evaluation of cross-lingual DocRE. We assess the cross-lingual transferability of existing DocRE models between English and Japanese, from which we observe challenges due to the complexity of document semantics. Our dataset will be publicly available[1].

## 2. Dataset Construction

**Task Definition.** For each document $D$ consisted of $n$ sentences $\mathcal{X}_D = x_1, x_2, \ldots, x_n$, entities within the document are given as $\mathcal{E}_D = \{e_1, e_2, \ldots, e_k\}$, where each entity $e_i \in \mathcal{E}_D$ is a collection of all its proper-noun mentions $e_i = \{m_1^i, m_2^i, \ldots, m_l^i\}$. A DocRE model is expected to extract all relation triples within the document in the form of $(e_h, r, e_t)$, where $e_h$ is the head entity, $e_t$ is the tail entity, and $r$ is a relation label chosen from a predefined set. Additionally, we also expect the model to perform *evidence retrieval*, where evidence for each relation prediction is provided at the sentence level. In other words, for a predicted triple $(e_h, r, e_t)$, the model should be able to return the evidence sentences $\mathcal{V}_{e_h,r,e_t} \subseteq \mathcal{X}_D$.

**Approach.** We explore ways to construct language resources for Japanese DocRE using existing English resources. To do so, we start by automatically building a dataset with cross-lingual transfer(Section 2.1). The approach has been reported successful in other IE tasks (Chen et al., 2023b; Hennig et al., 2023); If the transferred dataset portrays the characteristics of Japanese DocRE well, there is no need to recruit human annotators. However, we observe that the DocRE models trained with such a dataset err on raw Japanese text. Nevertheless, the model yields predictions of fair quality. We thus adopt the model trained on the transferred dataset as an intermediary tool to assist human annotation (Section 2.2).

[1]The dataset is available at https://github.com/YoumiMa/JacRED

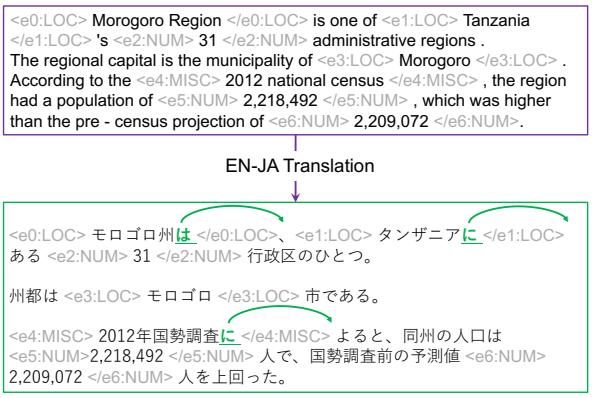


Figure 2: Transferring Re-DocRED from English into Japanese. We post-edit the translation to detach case markers from entity spans.

## 2.1. Automatic Construction

We build a Japanese version of Re-DocRED (Tan et al., 2022b). Re-DocRED revises DocRED (Yao et al., 2019), the first and most popular DocRE dataset collected from English Wikipedia.

**Translation and Annotation Projection.** We translate the complete train/dev/test splits of Re-DocRED into Japanese with the help of machine translators. As shown in Figure 2, XML tags are inserted around each entity. Documents are translated from English to Japanese with the tags so that entity spans are projected jointly during the translation process. Relations associated with can be thus directly inherited from the English dataset. This mark-then-translate method has been reported to work well for multiple structured prediction tasks (Chen et al., 2023b). We utilize DeepL to perform translation, as it enables translation while preserving XML tag markups[2].

**Post-processing for Case Markers.** Given the translation as in Figure 2, we recognize the necessity of post-editing due to the presence of case markers in entity spans. Case markers ("kaku-joshi" in Japanese) are special linguistic units attached to the end of nouns to indicate the relationship between words. A case marker only reveals the grammatical role but does not contribute to the semantics of the noun phrase it is attached to. For example, in entity span **<e0>** of the Japanese translation, a topic marker "は" following "モロゴロ州" (Morogoro Region) indicates the noun phrase to be the topic of this sentence. We detach case makers from the entity span with the Japanese morphological analyzer MeCab (Kudo et al., 2004)[3], removing tokens identified as particles at the end of each span. The obtained dataset is denoted as Re-DocRED$^{ja}$.

[2] https://api.deepl.com/v2/translate
[3] https://taku910.github.io/mecab/

**JA:** 堀 直宥(ほり なおさだ、寛文5年11月17日(1665年12月23日) - 正徳元年6月8日(1711年7月23日))は、江戸時代前期から中期の大名で、上総八幡藩第2代藩主。

**EN:** Naosada Hori (December 23, 1665 - July 23, 1711) was a feudal lord of the early to mid-Edo period, the second lord of the Joso Hachiman domain.

**missed triple**: (Naosada Hori, head of government, Joso Hachiman domain)

(a) Example of unextracted relations due to the topic shift of contents. The highlighted "藩主" is a Japanese historical term used from 1603 to 1912 meaning "lord".

**JA:** ザカリアーシュ・ヨージェフ(1924年3月25日 - 1971年11月22日)は、ハンガリー出身のサッカー選手、サッカー指導者。
1954年のFIFAワールドカップでは決勝戦を除く4試合にフル出場し準優勝に貢献した。

**EN:** Zakarias Yogev (March 25, 1924 - November 22, 1971) was a Hungarian soccer player and soccer coach.
(He) played in all but the final four games of the 1954 FIFA World Cup, contributing to the runners-up finish.

**missed triple**: (Zakarias Yogev, participant in, the 1954 FIFA World Cup)

(b) Example of unextracted relations due to the gap of surface structures. The subject of the second sentence is left out in Japanese.

Figure 3: Cases where the model trained on Re-DocRED$^{ja}$ failed to predict. Documents are shown as partial for better visibility. Note that English translations are provided only for reference, while predictions are actually done on Japanese texts.

**Limitations of Transferred Dataset.** When utilizing Re-DocRED$^{ja}$ as the training data and test bed, we witness DREEAM (Ma et al., 2023), the current state-of-the-art DocRE model, achieving an F1 score of 72.74 (cf. the same architecture scores 77.94 on the original Re-DocRED). However, when "real" Japanese documents from Japanese Wikipedia are fed into the model, we observe relation triples being left out in the predictions, with typical examples demonstrated in Figure 3. Two possible reasons can be raised to explain why the model trained on Re-DocRED$^{ja}$ fails: (1) **Topic Shift of Contents**: Re-DocRED$^{ja}$ cannot represent the real topic distribution of Japanese documents. Collected from English Wikipedia, Re-DocRED consists of contents that English speakers are concerned about, which do not necessarily match the interests of Japanese speakers. As in Figure 3a, Re-DocRED$^{ja}$ lacks documents about Japanese culture, preventing the DocRE models from being localized. (2) **Gap of Surface Structures**: The surface structures, i.e., how words are organized in the sentence, of Re-DocRED$^{ja}$ follow the logic of English, which is distinct from that of Japanese. Figure 3b showcases a typical example of how Japanese differs from English in surface structures regarding the omission of subjects. Re-DocRED$^{ja}$ thus cannot reproduce the surface structures of "real" Japanese, resulting in failures of the trained model.

## 2.2. Semi-Automatic Construction

Having observed drawbacks of Re-DocRED$^{ja}$, we postulate that human annotations are necessary to better depict Japanese DocRE. We thus involve human annotators in constructing a Japanese DocRE dataset, which we call JacRED. The annotation process consists of two phases: the entity mention annotation phase and the relation annotation phase. Both phases follow an edit-based scheme (Yao et al., 2019): Annotators only need to edit machine recommendations instead of listing all relation instances from scratch.

The quality of machine recommendation is crucial under the edit-based scheme: Poor recommendations require more edits, which will drastically increase the annotators' workload and affect the dataset's quality. The problem is recognized in DocRED as the *false-negative issue*, where too many relation instances are left out in the recommendations to be mended by human edits (Huang et al., 2022). We propose to mitigate this issue using Re-DocRED$^{ja}$, utilizing a model trained on Re-DocRED$^{ja}$ to recommend relation instances.

**Documents.** JacRED is built on top of the Japanese edition of Wikipedia. We clean up the dump and extract the opening text of each page as the document[4], with only those longer than 256 characters kept in our annotation pool.

**Annotators.** Given the complexity of the task, we recruit native Japanese speakers with expertise in annotating language resources instead of crowdsourcing[5]. The annotators first work individually on different data and then cross-check the worked annotations. The annotation tool is BRAT (Stenetorp et al., 2012) during both phases.

### 2.2.1. Entity Mention Annotation

The purpose of the entity annotation phase is twofold: (1) to obtain high-quality entity mention annotations for each document and (2) to filter out documents involved with few entities and relations.

**Entity Types.** We adopt the definition of IREX (Information Retrieval and Extraction Exercise, Sekine and Isahara (2000)) with 8 types, whose scope is similar to that of DocRED. A list of entity types is provided in Table 7 of Appendix A.

**Machine Recommendations.** We parse each document and obtain machine predictions of named entity mentions using KWJA (Ueda et al., 2023), a unified analyzer for Japanese.

**Document Filtering.** Another round of document filtering is performed based on the machine prediction to remove documents that are likely to contain few **cross-sentence** relations. To this end, we first link each mention to Wikidata entities (Vrandečić and Krötzsch, 2014). If an edge with label $r$ connects a certain entity pair $(e_h, e_t)$ in the knowledge base, we treat $(e_h, r, e_t)$ as an extractable relation triple from the document, following the distant-supervision assumption (Mintz et al., 2009). Only documents with more than 4 cross-sentence relations are preserved in the annotation pool. We employ mGENRE (De Cao et al., 2022) for entity linking and KGTK (Ilievski et al., 2020) for connectivity check.

**Human Edits.** We randomly select 2,000 documents from the annotation pool for human annotation. Human annotators review recommendations in each document, correcting wrongly predicted entity mentions and supplementing missed ones.

### 2.2.2. Relation Annotation

Relations and coreferences are annotated based on entities. Our approach differs from existing works in that (1) we define a smaller relation label set that covers a sufficient number of relation instances, and (2) we provide machine recommendations with a model trained with Re-DocRED$^{ja}$.

**Coreference Recommendations.** For each entity $e_i$, we treat all its mentions $\{m_1^i, \ldots, m_l^i\}$ as coreferences of each other. As introduced in the task definition, we only consider proper nouns as mentions while excluding the pronouns. Mentions linked to the same Wikidata entity are recommended as coreferences.

**Relation Types.** (Re-)DocRED's relation label set $\mathcal{R}$ contains 96 relation types. However, it is hard for annotators to comprehend such a large label set, which will eventually affect the annotation quality. We thus reduce the relation label set based on the following principles: (1) All relation categories defined in ERE (Song et al., 2015) should be covered; (2) Explicitly-defined inverse relation pairs, e.g., *has_part* and *part_of*, are merged into one; (3) Relations frequently appearing in Re-DocRED are preserved as much as possible. This results in a label set $\mathcal{R}'$ of 28 relations covering over 88% relation instances in Re-DocRED.

[4] 2023-01-01 dump at https://dumps.wikimedia.org/jawiki/

[5] Measures including the Inter Annotator Agreements (IAA) are thus not reported in this paper.

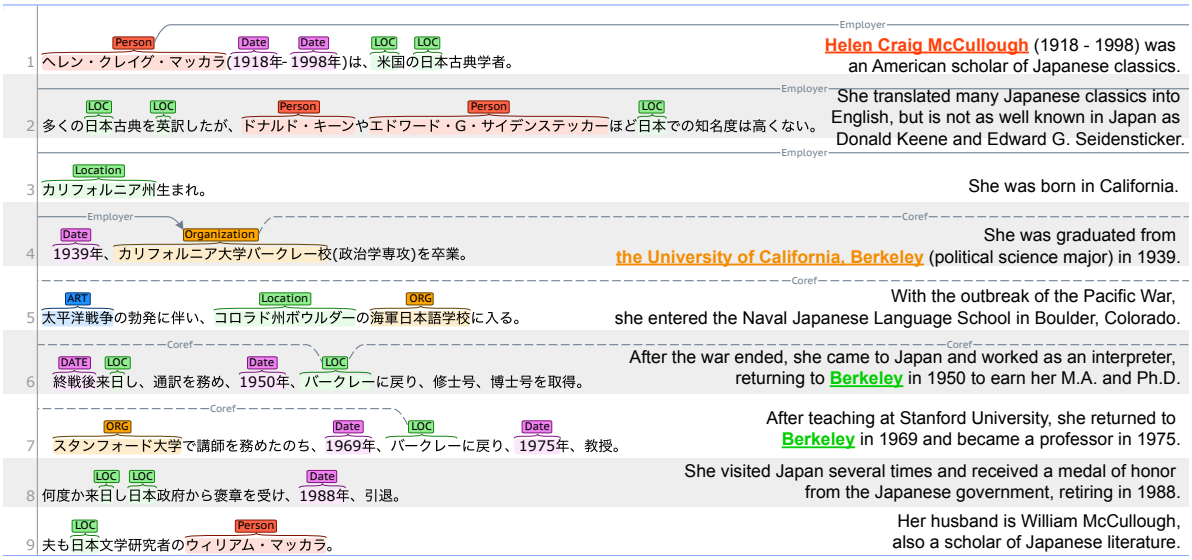


Figure 4: Interface for relation annotation. English translations are provided on the right for reference. In this example, the annotator decides whether (Helen Craig McCullough, Employer, the University of California, Berkeley) holds or not. Entity mentions connected with *Coref* are coreferences of each other.

**Relation Recommendations.** We project the label set of Japanese Re-DocRED from $\mathcal{R}$ to $\mathcal{R}'$ and retrain a DREEAM model. Predictions of the model are employed as machine-recommended relations. We expect our recommendations to be more accurate than those in previous works obtained from knowledge base queries, primarily due to two factors: (1) Wikidata only stores a limited number of relation facts, while a model can, in principle, assign relation(s) to each entity pair in the document; (2) Relation facts in Wikidata are independent of the document's content, while model predictions are contextually sensitive. A quantitative comparison of recommendations from the model trained on Re-DocRED$^{ja}$ and those from querying Wikidata can be found in Section 3.2.

**Human Edits.** Coreferences and relations are revised during human annotation. For coreferences, human annotators remove irrelevant mentions and supplement missed mentions for each entity. For relations, human annotators first examine the existence of each recommended relation. As showcased in Figure 4, a pair of mentions $m_i^h, m_i^t$, representing entity $e_h, e_t$ respectively, along with their relation $r$ is shown in the interface. If annotators consider relation triple $(e_h, r, e_t)$ as true, they need to provide the evidence sentence $\mathcal{V}_{e_h,r,e_t}$ within the document[6]; Otherwise, the triple should be deleted from the dataset. Finally, the annotators supply missing relation triples and evidence sentences with their best effort.

[6]Sentences where mention $m_i^h$ and $m_i^t$ resides are treated as evidence by default. Only evidence sentences other than those need to be provided.

**Post-processing.** Among all 28 relation types, 7 have inverse relations defined in Wikidata. We automatically augment triples of inversed relation types after human edits. For example, if triple $(e_h, part\_of, e_t)$ is present in the revised annotation and relation type $part\_of$ is an inversion of $has\_part$, a new triple $(e_t, has\_part, e_h)$ will be automatically added into the annotation. JacRED thus includes 35 relation types eventually. A detailed list of relation types is provided in Table 8 of Appendix B.

## 3. Dataset Analysis

This section reports the analysis results of JacRED to provide a deeper understanding of the collected dataset. Firstly, we compare the statistics of JacRED against (Re-)DocRED. The comparison suggests that JacRED combines the advantages of DocRED and Re-DocRED (Section 3.1). Next, we calculate the number of edits human annotators made before reaching the final annotations. We observe that significantly more edit steps would be necessary if the human annotation started from machine recommendations suggested by knowledge base queries (Section 3.2).

### 3.1. Detailed Statistics

Table 2 details the agreements and differences between (Re-)DocRED and JacRED.

**Document Complexity.** As for document length, JacRED shares a similar scale with (Re-)DocRED at both token and sentence levels. On

| | DocRED | Re-DocRED | JacRED |
|---|---|---|---|
| # Sentences | 7.98 | 7.98 | 8.39 |
| # Entities | 19.51 | 19.45 | 17.87 |
| # Relations | 12.45 | 29.77 | 21.12 |
| # Evidences | 1.60 | 0.88 | 1.67 |

Table 2: Comparison of (Re-)DocRED and JacRED. Values are average for each document.

one hand, documents in JacRED contain more relation instances than DocRED on average, implying that the false negative issue is mitigated in JacRED compared to DocRED. On the other hand, documents in JacRED contain fewer relation instances than in Re-DocRED. One possible reason is our re-definition of relation types, where symmetric relation types are merged into one as they represent the same knowledge.

**Evidence Annotation.** Re-DocRED revises DocRED to alleviate the false negative issue by supplying missed relation instances. However, evidence sentences for those supplied instances are not included in Re-DocRED. In contrast, we collect human-annotated evidence sentences during the relation annotation phase. JacRED thus better portrays the correlation between relation and evidence sentences than Re-DocRED.

### 3.2. Number of Human Edits

We quantify the distance between machine recommendations and human annotations of relation instances. To this end, we compare machine recommendations against final human annotations to see how many edits have been made. Specifically, we randomly sample 400 documents from JacRED and calculate the number of recommendations being deleted/substituted/supplied as in Table 3.

**Human Annotations v.s. Machine Recommendations.** We observe that more than 20% of machine recommendations (1,490 out of 6,500) were regarded as inappropriate, whose relation labels were deleted or substituted by human annotators. The human annotators also supplied another 2,740 relation instances, taking up more than 40% of the recommendations. We thus conclude that DocRE models trained on the automatically constructed dataset still lag behind human performance considerably, suggesting the importance of a manually collected dataset.

**Cross-Lingual Transfer v.s. Knowledge Base Queries.** We further measure the distance of human annotations from relations recommended by querying Wikidata, a de-facto method used in previous works (Yao et al., 2019; Cheng et al., 2021). Compared with model predictions, Wikidata provides only half as many recommendations: To reach the human annotations, 50% (1,572 out of 3,200) of the recommendations need to be revised, with another 200% instances to be added. In total, it takes 7,805 steps to align Wikidata recommendations with the final annotation, while only 4,230 steps are needed when employing cross-lingual transfer. These statistics reveal the usefulness of Re-DocRED$^{ja}$ in reducing human efforts.

## 4. Experiments

**Purposes.** We employ JacRED as a benchmark to examine the capability of existing DocRE models. Our major concerns are: (1) How well can existing DocRE models perform Japanese DocRE? (2) How different can a DocRE model perform when trained on Re-DocRED$^{ja}$ and JacRED? Additionally, we evaluate the cross-lingual transferability of existing DocRE models with JacRED.

**Settings.** We split JacRED into train/dev/test sets with 1400/300/300 documents. Models are trained and evaluated on a single Tesla V100 16GB GPU. For evaluation, we follow previous works to compute the micro-averaged F1 scores for relations and evidence sentences (Yao et al., 2019). Additionally, we compute **Rel F1 Ign**, a variant of F1, where relation instances seen in the training set are ignored during evaluation. Average scores of 5 runs initialized with different random seeds are reported throughout this paper.

### 4.1. Models Trained on JacRED

We measure the performance of existing models when supervised by the training split of JacRED. Specifically, we train and evaluate 4 popular models on top of *tohoku-nlp/bert-base-japanese-v2* available on Huggingface[7], with results summarized in Table 4. Among these models, DREEAM is the current state-of-the-art model on (Re-)DocRED for extracting both relations and evidence sentences. We also evaluate the performance of LLM with in-context learning.

**JacRED introduces extra challenges beyond those in Re-DocRED.** In Table 4, all DocRE models score above 60 on Relation F1. Although acceptable, the performance of each model is worse than their equivalents trained on Re-DocRED, with a gap of 10 F1 points (cf. Table 6). The result suggests potential challenges in

[7] https://huggingface.co/tohoku-nlp/bert-base-japanese-v2

| | # Recommendations | # Deletions | # Substitutions | # Supplements |
|---|---|---|---|---|
| Cross-lingual Transfer | 6,500 | 1,266 | 224 | 2,740 |
| Knowledge Base Queries | 3,200 | 1,459 | 113 | 6,233 |

Table 3: Number of relation instances automatically recommended and how they should be revised to reach the final human annotations.

| | Dev Set | | | Test Set | | |
|---|---|---|---|---|---|---|
| | Rel F1 | Rel F1 Ign | Evi F1 | Rel F1 | Rel F1 Ign | Evi F1 |
| ATLOP (Zhou et al., 2021) | 66.53 | 65.21 | – | 68.04 | 66.80 | – |
| DocuNet (Zhang et al., 2021) | 66.67 | 65.37 | – | 67.66 | 66.47 | – |
| KD-DocRE (Tan et al., 2022a) | 67.12 | 65.70 | – | 68.29 | 66.99 | – |
| DREEAM (Ma et al., 2023) | **67.34** | **65.90** | **61.46** | **68.73** | **67.40** | **62.11** |
| *gpt-3.5-turbo-instruct* (Ouyang et al., 2022) | 13.46 | 12.84 | – | 13.17 | 12.90 | – |
| *gpt-4* (OpenAI, 2023) | 24.17 | 23.63 | – | 27.45 | 26.99 | – |

Table 4: Models' performance on the development and test set of JacRED, with best scores **bolded**.

JacRED that are absent from Re-DocRED, possibly due to the characteristics of the Japanese language, such as the omission of subjects. Addressing such characteristics may be essential to better tackle Japanese DocRE.

**In-context learning of LLMs on JacRED is non-trivial.** Apart from models specially designed for DocRE, we also evaluate how LLMs can tackle the task via in-context learning. Specifically, we pre-define the relation label set and include a pair of (document, relations) in the prompt to guide the LLM in conducting DocRE. In our experiments, we utilize models provided by OpenAI, namely (1) the instructed version of GPT-3.5 accessed via the API key *gpt-3.5-turbo-instruct* (Ouyang et al., 2022) and (2) GPT-4 accessed via the API key *gpt-4*[8]. As in the last two rows of Table 4, GPT-3.5 exhibited much lower performance than the DocRE models. GPT-4 improved over GPT-3.5 but still lagged behind the supervised DocRE models. Similar insights have been provided by Wadhwa et al. (2023), where in-context learning of DocRE could not be conducted due to the length restriction of the prompt. We succeeded in instructing LLM to conduct DocRE, while the performance is limited. The experiment results thus highlight the challenge of DocRE as a task that LLMs cannot easily tackle.

## 4.2. Models Trained on Transferred Re-DocRED

Section 2.1 has mentioned limitations in the dataset automatically constructed from cross-lingual transfer. Specifically, we showcased how DocRE models trained on such a dataset fail to extract relation triples from raw Japanese documents. Here, we quantify the performance gap between a model trained on the automatically constructed dataset (Re-DocRED$^{ja}$) and the human-annotated dataset with machine assistance (JacRED). The test set of JacRED is adopted as the benchmark, with results shown in Table 5.

| | Relation | | |
|---|---|---|---|
| Training Data | P | R | F1 |
| JacRED (1,400) | **64.76** | **73.29** | **68.73** |
| Re-DocRED$^{ja}$ (3,053) | 56.14 | 53.67 | 54.87 |
| Re-DocRED$^{ja}$ (1,400) | 55.52 | 51.77 | 53.56 |

Table 5: Precision (P), Recall (R), and F1 scores of DREEAM trained on different data, evaluated on the test set of JacRED. The number of documents in each set is shown in parentheses.

**Models trained on Re-DocRED$^{ja}$ suffer from low recalls.** From Table 5, we witness that DREEAM trained on Re-DocRED$^{ja}$ underperforms its equivalent trained on JacRED. Taking a closer look at the scores, we find the gap in recalls (73.29 v.s. 53.67) is more significant than that in precisions (64.76 v.s. 56.14). The result corresponds to our observation in Section 2.1 that models trained on the transferred dataset cannot identify some relation instances due to the limitation of texts translated from English.

**The gap between models trained on Re-DocRED$^{ja}$ and JacRED is evident under the same setting.** We further train DREEAM on Re-DocRED$^{ja}$ with 1,400 documents, aligned with the number of documents in JacRED. The F1 score drops from 54.87 to 53.56, lagging behind that of the model trained on JacRED with a gap of 15 F1 points. The results indicate that JacRED provides

[8]Details of the prompt is provided in Figure 5 of Appendix C.

| | Rel (*tgt*) | | | Rel (*src*) |
|---|---|---|---|---|
| **Model** | **P** | **R** | **F1** | **F1** |
| **(a) *en.* → *ja.*** | | | | |
| ATLOP | **60.59** | 31.91 | 41.76 | 74.82 |
| DocuNet | 60.44 | 34.50 | 43.92 | 75.02 |
| KD-DocRE | 58.83 | **36.67** | 45.14 | 75.72 |
| DREEAM | 60.07 | 36.36 | **45.29** | **77.22** |
| **(b) *ja.* → *en.*** | | | | |
| ATLOP | 53.13 | 48.70 | 50.72 | 64.25 |
| DocuNet | 52.69 | 45.85 | 49.03 | 64.64 |
| KD-DocRE | **54.22** | 50.12 | 52.09 | 65.42 |
| DREEAM | 51.88 | **53.05** | **52.45** | **65.90** |

Table 6: Cross-lingual performance on the test set of JacRED (*ja.*) and Re-DocRED (*en.*) of models with mBERT as the encoder.

better supervision than Re-DocRED[ja].

### 4.3. Crosslingual DocRE

JacRED also enables the evaluation of cross-lingual DocRE. Although DocRE datasets have been collected in Chinese (Cheng et al., 2021) and Korean (Yang et al., 2023), they lay in different domains than (Re-)DocRED. In contrast, JacRED is collected from Wikipedia following a pipeline similar to DocRED. The domain and label sets of JacRED and (Re-)DocRED thus match each other, enabling the evaluation of cross-lingual DocRE. Here, we take the first attempt to measure the cross-lingual transferability of existing models using Re-DocRED and JacRED.

Specifically, we train models on the training set in one language and evaluate them on the test set in another. The relation label set of Re-DocRED is projected onto JacRED using the same method as in Section 2.2. To ensure the multilingualism of trained models, we adopt multilingual BERT (mBERT, Devlin et al. (2019)) as the encoder. Evaluation results are shown in Table 6.

**Cross-lingual performance of existing models is limited.** All models exhibited a decreased accuracy in the target language. Different from sentence-level tasks, DocRE requires not only an understanding of individual sentences but also inter-sentence semantics within the whole document, which improves the difficulty of building cross-lingual models. This may offer a potential explanation as to why translation-based cross-lingual transfer is ineffective for DocRE, despite its successful application in sentence-level RE and OpenIE (Kolluru et al., 2022; Hennig et al., 2023).

## 5. Related Work

**DocRE corpora in English.** The most well-known definition of DocRE was proposed by Yao et al. (2019), along with a dataset collected from English Wikipedia named DocRED. While two document-level relation extraction datasets, namely CDR (Li et al., 2016) and GDA (Wu et al., 2019), have been proposed ahead of DocRED, they were collected in the biomedical domain, thus unsuitable for developing general-purpose DocRE models. DocRED suffers from the false negative issue where a considerable amount of relation instances are absent from the ground-truth annotations (Huang et al., 2022; Xie et al., 2022; Tan et al., 2022b). Huang et al. (2022) randomly selected 96 documents from DocRED and relabeled them from scratch, while Tan et al. (2022b) revised the whole dataset as Re-DocRED with machine assistance. This work follows a machine-assisted annotation process as DocRED and Re-DocRED while paying extra attention to providing better machine recommendations with the model trained on a dataset transferred from Re-DocRED.

**DocRE corpora in other languages.** Cheng et al. (2021) constructed HacRED from Chinese DBpedia to promote relation extraction from complex contexts. Yang et al. (2023) focused on Korean historical RE research and collected HistRED from a travel diary written between the 16th and 19th centuries. These datasets were collected independently from DocRED with distinct domains and label sets. Apart from these studies, Cheng et al. (2022) released a system for medical relation extraction on Japanese documents, while the dataset is not publicly available. In this work, we explore how existing resources can help construct DocRE resources in other languages. We share the insights that models trained under cross-lingual transfer techniques are not ready for practical use. However, they serve as good assistants for aiding human annotations.

**Cross-lingual transfer for structured predictions.** Several works have adopted translation-based cross-lingual transfer approaches to solve cross-lingual and multi-lingual structured prediction tasks (Faruqui and Kumar, 2015; Kolluru et al., 2022). More recently, Hennig et al. (2023) constructed MultiTACRED, a multilingual version of TACRED (Victor Zhong et al., 2018), using similar approaches as ours. They confirmed the dataset's quality to be high enough even without human modifications. Our work examines the approach's usefulness in the literature of DocRE and reports its shortcomings. Unlike other sentence-level IE tasks, DocRE involves understanding not only sin-

gle sentences but also the whole document, improving the difficulty of cross-lingual transfer.

## 6. Conclusion

This work publishes JacRED, the first benchmark for general-purpose Japanese DocRE. In the process of building JacRED, we explore how to utilize existing English DocRE resources to construct resources for other languages, using Japanese as the representative. Starting from constructing a dataset by translation-based cross-lingual transfer, we have shown how and why such a dataset is not ready for practical use. Nevertheless, models trained on the dataset can replace existing approaches, i.e., knowledge base queries, to provide better recommendations for human annotation. Our insights can benefit the development of DocRE resources for other languages. Benchmarking with JacRED portrays the challenge of not only Japanese but also cross-lingual DocRE.

In the future, we plan to utilize models trained on JacRED to help downstream tasks such as question answering and reading comprehension.

## 7. Ethics Statement

In this work, we collected a dataset from Wikipedia, whose text content can be used under the terms of the CC-BY-SA[9]. We thus presume that no copyright issues are involved in constructing and publishing our dataset.

**Automatic Annotations.** For the machine translator, we adopted DeepL API at the cost of 2,500 JPY (approx. 16$) per 1 million characters. For the LLM, we tested with the instructed version of GPT-3.5 provided by OpenAI at the cost of $0.0015 per 1 thousand tokens for input and $0.002 per 1 thousand tokens for output. For existing DocRE models, all resources we adopted are publicly available and free of charge.

**Human Annotations.** Before the annotation, we arranged meetings in advance to (1) explain the purpose of collecting the dataset and (2) adjust the workload. The annotators understand and agree that their work will be used to train neural networks. For both the entity and relation annotation phases, we explained the purpose of building the dataset and provided a detailed annotation guideline. During the annotation, we frequently discussed with the annotators how to handle irregular cases and adjust the guidelines when necessary. 7 annotators are involved in the entity mention annotation phase, and 6 annotators are involved in the relation annotation phase. Each annotator is paid 5,000 JPY (approx. 30$) per hour, which is higher than the standard salary in Japan.

[9] https://en.wikipedia.org/wiki/Wikipedia:Text_of_the_Creative_Commons_Attribution-ShareAlike_4.0_International_License

## 8. Acknowledgements

This paper is based on results obtained from a project, JPNP18002, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

## 9. Bibliographical References

Haotian Chen, Bingsheng Chen, and Xiangdong Zhou. 2023a. Did the models understand documents? benchmarking models for language understanding in document-level relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6418–6435, Toronto, Canada. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023b. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. 2022. JaMIE: A pipeline Japanese medical information extraction system with novel relation annotation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3724–3731, Marseille, France. European Language Resources Association.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online. Association for Computational Linguistics.

Barry Chiswick and Paul Miller. 2004. Linguistic distance: A quantitative measure of the distance between english and other languages. IZA Discussion Papers 1246, Institute of Labor Economics (IZA).

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada. Association for Computational Linguistics.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in DocRED. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6241–6252, Dublin, Ireland. Association for Computational Linguistics.

Filip Ilievski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Ronpeng Li, Jun Liu, Amandeep Singh, Daniel Schwabe, and Pedro Szekely. 2020. KGTK: A toolkit for large knowledge graph manipulation and analysis. In *International Semantic Web Conference*, pages 278–293. Springer.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel 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, Suntec, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Satoshi Sekine and Hitoshi Isahara. 2000. IREX: IR & IE evaluation project in Japanese. In

*Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

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

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy. Association for Computational Linguistics.

Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2023. KWJA: A unified Japanese analyzer based on foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 538–548, Toronto, Canada. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *Research in Computational Molecular Biology*, pages 272–284, Cham. Springer International Publishing.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland. Association for Computational Linguistics.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581, Vancouver, Canada. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International*

*Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

## 10. Language Resource References

Victor Zhong et al. 2018. *TAC Relation Extraction Dataset*. The Stanford NLP Group. Linguistic Data Consortium, The Stanford NLP Group resources, 1.0, ISLRN 927-859-759-915-2.

| (Re-)DocRED (6) | JacRED (8) |
|---|---|
| PERSON | PERSON |
| ORGANIZATION | ORGANIZATION |
| LOCATION | LOCATION |
| TIME | ARTIFACT |
| NUM | TIME |
| MISC | DATE |
| | PERCENT |
| | MONEY |

Table 7: Comparison of entity types of existing dataset and our proposed dataset. The total number of entity types is indicated in the parenthesis following each dataset.

## Appendix A: Entity Label Types

In this section, we list all entity types in JacRED as in Table 7, together with those defined in (Re-)DocRED.

## Appendix B: Relation Label Types

In this section, we list all relation types included in JacRED as in Table 8.

## Appendix C: Prompt for In-Context Learning

We showcase the prompt used for the in-context learning of LLM in Figure 5. In previous work where LLM are utilized for relation extraction (Wadhwa et al., 2023; Li et al., 2023), the prompt has been designed to return all relation triples within a document. However, it is hard to identify all relation triples across a document at once. Furthermore, most supervised approaches tackle DocRE by classifying relation types entity-pair wise (Zhou et al., 2021; Xie et al., 2022; Tan et al., 2022a; Ma et al., 2023). We thus design prompts to reduce the task complexity by querying one relation type for each API call[10]. By using our prompt, GPT-3.5 yields better performance than reported in existing works.

[10] In early experiments, we evaluated the performance when querying: 1) one relation type; 2) all relation types of one entity pair; 3) one relation type of one entity pair during each API call, where 2) yields good performance at a low cost.

| ERE Category | JacRED Type | ID |
|---|---|---|
| Physical | Capital | P36 |
| | CapitalOf | P1376 |
| | AdministrativeLocation | P131 |
| | Location | P276 |
| | WorkLocation | P937 |
| General Affiliation | CountryOfCitizenship | P27 |
| | DateOfBirth | P569 |
| | DateOfDeath | P570 |
| | PlaceOfBirth | P19 |
| | PlaceOfDeath | P20 |
| | Follows | P155 |
| | FollowedBy | P156 |
| Personal-Social | Child | P40 |
| | Sibling | P3373 |
| | Spouse | P26 |
| | ParticipantIn | P1344 |
| | Participant | P710 |
| Part-Whole | MemberOf | P463 |
| | HasPart | P527 |
| | PartsOf | P361 |
| Organization Affiliation | HeadOfGovernment | P6 |
| | OwnedBy | P127 |
| | OwnerOf | P1830 |
| | FoundedBy | P112 |
| | Employer | P108 |
| | Operator | P137 |
| | ItemOperated | P121 |
| | EducatedAt | P69 |
| Others (*) | AwardReceived | P166 |
| | Creator | P170 |
| | Performer | P175 |
| | Published | P123 |
| | PresentInWork | P1441 |
| | Characters | P674 |
| | Platform | P400 |

Table 8: Relation types included in our proposed dataset. Column **ID** shows the Wikidata property ID linked to each relation type. The last category **Others** includes relation types undefined in ERE type.

Perform Document-level Relation Extraction task. Given a context and an entity list, identify all entity pairs with relation type {located in the administrative territorial entity} in the context. Note that only a few entity pairs hold relations. Please return entity pairs as {head, tail} and make sure they follow the relation definition:

located in the administrative territorial entity: {head} is located in the administrative territorial entity {tail}.

###

Context: 東京・板橋出身。

Entity List: 東京||板橋

Extracted Entity Pairs: {板橋, 東京}

###

Context: 南都六宗(なんとろくしゅう、なんとりくしゅう)とは、奈良時代、平城京を中心に栄えた日本仏教の6つの宗派の総称。三論宗(さんろんしゅう、中論・十二門論・百論)-華厳宗や真言宗に影響を与えた成実宗(じょうじつしゅう、成実論)-三論宗の付宗(寓宗)法相宗(ほっそうしゅう、唯識)倶舎宗(くしゃしゅう、説一切有部)-法相宗の付宗(寓宗)華厳宗(けごんしゅう、華厳経)律宗(りっしゅう、四分律)-真言律宗等が生まれたなお、奈良時代当時から「南都六宗」と呼ばれていたわけではなく、平安時代以降平安京を中心に栄えた「平安二宗」(天台宗・真言宗)に対する呼び名である。

Entity List: 奈良時代||平安時代||平城京||日本||平安京||平安

Extracted Entity Pairs: {平安京, 日本}

###

(examples)

###

Context: アンソニー世界を駆ける(アンソニーせかいをかける)は、アメリカ合衆国のCNNで放送されているテレビ番組。2013年4月から放送を開始した。エミー賞を4回受賞、また、脚本賞、音響賞、編集賞、撮影賞に11回ノミネートされている。また2013年にはアメリカのテレビ・ラジオ・ウェブサイトの優れた放送作品に贈られるピーボディ賞を受賞した。自ら料理人であり、ノンフィクション「キッチン・コンフィデンシャル」の著者でもあるアンソニー・ボーディンが世界の津々浦々を旅し、あまり知られていない地域の景観、風俗、食材、料理などを紹介する。

Entity List: アメリカ合衆国||アンソニー世界を駆ける||CNN||2013年4月||エミー賞||2013年||ピーボディ賞||キッチン・コンフィデンシャル||アンソニー・ボーディン

Extracted Entity Pairs:

Figure 5: An example of the prompt used for the in-context learning of GPT-3.5 and GPT-4.