Towards Realistic Few-Shot Relation Extraction: A New Meta Dataset and Evaluation

Fahmida Alam[1], Md Asiful Islam[1], Robert Vacareanu[1,2], Mihai Surdeanu[1]
[1]University of Arizona, Tucson, USA
[2]Technical University of Cluj-Napoca, Cluj-Napoca, Romania
{fahmidaalam, asifulislam, rvacareanu, msurdeanu}@arizona.edu

Abstract

We introduce a meta dataset for few-shot relation extraction, which includes two datasets derived from existing supervised relation extraction datasets – NYT29 (Takanobu et al., 2019; Nayak and Ng, 2020) and WIKIDATA (Sorokin and Gurevych, 2017) – as well as a few-shot form of the TACRED dataset (Sabo et al., 2021). Importantly, all these few-shot datasets were generated under realistic assumptions such as: the test relations are different from any relations a model might have seen before, limited training data, and a preponderance of candidate relation mentions that do not correspond to any of the relations of interest. Using this large resource, we conduct a comprehensive evaluation of six recent few-shot relation extraction methods, and observe that no method comes out as a clear winner. Further, the overall performance on this task is low, indicating substantial need for future research. We release all versions of the data, i.e., both supervised and few-shot, for future research.[1]

Keywords: relation extraction, few-shot learning, evaluation

1. Introduction

Information Extraction (IE) plays a pivotal role in Natural Language Processing (NLP). IE is fundamental to many NLP tasks such as question answering, event extraction, knowledge base population, etc. Relation Extraction (RE) is a sub-task of IE with the focus of identifying entities and their semantic relations in a given text, enabling the extraction of structured information from unstructured data. For instance, in the sentence "John Doe was born in New York City", Relation Extraction can transform this into a structured tuple such as $\rightarrow$ (John Doe, `born in`, New York City), indicating the inherent relation between the person, action, and location.

Many supervised methods have been proposed to address the relation extraction task (Soares et al., 2019; Zhang et al., 2018; Wang et al., 2016; Miwa and Bansal, 2016, inter alia). However, a traditional supervised machine learning (ML) setup is not always realistic for RE due to the large amount of training data required. This setup is mostly incompatible with real-world RE scenarios such as pandemic response or intelligence, in which RE models must be developed and deployed quickly with minimal supervision.

Considering this task setup, a realistic choice for solving this problem is few-shot learning (FSL) and its RE equivalent, few-shot relation extraction (FSRE), in which (a) each relation class is associated with a very small number of examples (typically 1 or 5), and the relation classes in the testing partition are different from any relations a model might have seen before. While several FSRE datasets and methods have been proposed recently (see Related Work), this subfield of NLP is still poorly understood due to a lack of realistic datasets and rigorous evaluations. This observation has motivated this work, in which we introduce a meta dataset for the task as well as a meaningful evaluation of multiple FSRE methods on this data. The key contributions of our work are:

(a) We develop a meta dataset for FSRE, which includes three datasets: one based on NYT29 (Takanobu et al., 2019; Nayak and Ng, 2020), one based on WIKIDATA (Sorokin and Gurevych, 2017), and lastly the few-shot variant of TACRED proposed by (Sabo et al., 2021). All these datasets were converted into realistic few-shot variants using the procedure detailed in § 3.4. This procedure guarantees a setup that is aligned with real-world applications, e.g., the test relations are different from any relations available in a background dataset, limited training data, preponderance of candidate relation mentions that do not correspond to any of the relations of interest, etc.

(b) We conduct a comprehensive evaluation of six recent FSRE methods using this meta dataset. Our evaluation reveals that none of the models emerged as a definitive winner. Furthermore, the overall performance of the best models was notably low, indicating the substantial need for future research. Our

[1]Datasets and additional resources are available at: `https://github.com/clulab/releases/tree/master/lrec2024-realistic-fewshot-meta-dataset`

LREC-COLING 2024, pages 16592–16606

datasets will contribute as an invaluable resource for this future research.

2. Related Work

2.1. Methods

Historically, relation extraction approaches can be categorized as either rule-based or relying on statistical models. In the past decade, the latter category has been dominated by neural-based methods. More recently, hybrid directions have emerged, aiming to combine the advantages of both. We delve deeper into each of these directions.

2.1.1. Rule-based Methods

Prior to the widespread adoption of statistical machine learning, rule-based approaches enjoyed a period of prominence. These methods typically involve the acquisition of rules representative of specific relations. For example, the rule `[ne=PER]+ <nsubj born >nmod_in [ne=LOC]+` captures a syntactic pattern for the `born_in` relation, where the pattern matches if the underlying constraints are satisfied: a named entity labeled as person is connected to the word `born` with a `nominal subject` dependency, and the same word `born` is further connected to a named entity labeled as location with a `nominal modifier` dependency. For example, this pattern will match the sentence: *John Doe was born in New York City.* A match of such rules is then interpreted as an indication that the two entities participate in the corresponding relation.

In (Hearst, 1992), the authors propose a set of handwritten rules to extract words satisfying the hyponymy relation. Subsequently, efforts were directed toward automating the learning of such patterns (Riloff, 1993, 1996; Riloff and Jones, 1999) with and without supervision. (Gupta and Manning, 2014) improves automatic pattern learning by allowing soft matching in the form of predicting the labels on unlabeled entities.

Another prominent line of work for rule-based methods is that of casting the pattern learning problem as a graph-based problem and leveraging graph-based algorithms (Kozareva et al., 2008; Vacareanu et al., 2022a).

2.1.2. Neural-based Methods

The adoption of neural-based methods has grown significantly due to their high performance, making them the de facto approach for relation extraction tasks today. Many underlying architectures were proposed for relation extraction, such as ones based on CNNs (Zeng et al., 2014; Nguyen and Grishman, 2015), RNNs (Zhang and Wang, 2015), LSTMs (Zhang et al., 2017), or, more recently, Transformers (Vaswani et al., 2017; Joshi et al., 2019). These approaches typically operate end-to-end and are built on top of pre-trained embeddings, either static (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016) or contextual (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2019).

A more recent direction has been translating the relation extraction task into a different NLP task to leverage more training data (Chen et al., 2022). For example, relation extraction can be cast as an entailment problem (Sainz et al., 2021; Rahimi and Surdeanu, 2023), or as summarization (Lu et al., 2022).

A distinctive direction emerged in the last years, attempting to combine the advantages of both rule-based systems and neural-based systems. For example, Vacareanu et al. (2022b) teaches a neural network to generate rules for RE. Other directions aiming to improve the explainability of the resulting model include: (i) learning an explainability classifier jointly with the RE model to ensure faithfulness of explanations (Tang and Surdeanu, 2021, 2023), or (ii) learning a neural "soft" (or semantic) matcher to improve the rules' recall (Zhou et al., 2020).

2.2. Datasets and Methods for Few-Shot Relation Extraction

A key contribution to the RE space is the creation of datasets that support the development of new RE approaches. A recent survey (Bassignana and Plank, 2022) categorized popular relation classification datasets based on their data sources into three main categories: (i) News and Web, (ii) Scientific Publications, and (iii) Wikipedia, totaling 17 datasets. We refer the reader to this survey for more details.

An important and realistic setting for this task is few-shot relation extraction (FSRE), where only a small number of training examples are available for each relation class to be learned. Notably, only three datasets are available in a few-shot format (Bassignana and Plank, 2022): FewRel (Han et al., 2018), FewRel 2.0 (Gao et al., 2019), and few-shot TACRED (Sabo et al., 2021).

The FewRel dataset, containing 70,000 sentences covering 100 relations from Wikipedia, is created by identifying relation mentions through distant supervision; noise is subsequently filtered by crowd-workers (Han et al., 2018). Later on, the FewRel 2.0 dataset (Gao et al., 2019), an extension of the original FewRel dataset (Han et al., 2018), introduced a new test set in a distinct domain and included the option of a NOTA (None of the Above) relation.

Sabo et al. (2021) argues that FewRel provides an unrealistic benchmark due to its uniform relation distribution and the prevalence of proper nouns as entities. Although FewRel 2.0 tried to amend it using an updated episode sampling procedure, the evaluation setup is still notably unrealistic (Sabo et al., 2021). As a solution, Sabo et al. (2021) converted the supervised TACRED dataset (Zhang et al., 2017) into a few-shot TACRED variant by applying realistic episode sampling. Concretely, the episode in an FSRE evaluation should be selected in a way that follows all the criteria (a–f) we mention in Section 3.4. To develop our other few-shot datasets, i.e., NYT29 and WIKIDATA, we followed a similar strategy (see § 3.4 and 3.5).

Nevertheless, despite the unquestionable contribution of such datasets to the RE field, we observed a lack of consistency in the results observed in the various proposed evaluations. For example, some methods evaluated on FewRel attained an accuracy of 93.9%, surpassing human-level performance at 92.2% (Soares et al., 2019). While FewRel 2.0 yields lower results, i.e., the best method achieved 80.3% (Gao et al., 2019), they are still remarkably high, given the challenging nature of the task.

Further, (Sabo et al., 2021) evaluated their MNAV model (which was state-of-art at the time) on FewRel 2.0 and achieved an F1 score of approximately 78% for 5-way 1-shot and 80% for 5-way 5-shot, whereas the best results on TACRED are much lower: the F1 score is 12.4% for 5-way 1-shot and 30.0% for 5-way 5-shot. These differences are caused by differences in how the datasets are constructed, which impacts consistent analyses of the proposed methods. To remedy this issue, we propose a meta dataset for few-shot RE that includes three datasets that are constructed using the same realistic procedure and capture multiple important phenomena. This allows us to rigorously evaluate multiple approaches for few-shot RE as shown in § 5.2.

3. Dataset Construction Process

We detail next our first key contribution: the construction of a meta dataset for FSRE, which combines two new FSRE datasets and a third existing one.

3.1. Data Sources

We leverage three existing *supervised* datasets for RE to serve as our starting point. These datasets cover a diverse set of domains: NYT29, WIKIDATA, and TACRED.

NYT29: The NYT29 dataset originates from the New York Times corpus, which comprises a collection of more than 1.8 million articles authored and released by the New York Times between January 1, 1987, and June 19, 2007, with article metadata provided by the New York Times Newsroom (Sandhaus, 2008). This dataset was annotated with relations from Freebase using distant supervision by Riedel et al. (2010). Depending on how many relation classes are kept, this original dataset has multiple versions, e.g., "NYT10," "NYT11," and "NYT29" (Takanobu et al., 2019; Nayak and Ng, 2020). Our work relies on the latter version, which contains 29 distinct relations (e.g.,`/people/person/place_lived`), and it covers a wide range of topics, news events, and perspectives.

WIKIDATA: The WIKIDATA dataset is a subset of Wikipedia, wherein articles have been marked with Wikidata relations using distant supervision (Sorokin and Gurevych, 2017). This corpus encompasses two primary types of annotations: entities and relations. Entity annotations are derived from Wikipedia article links. Each link has been converted to a Wikidata identifier using the mappings from the Wikidata itself. Additional entities are recognized using a named entity recognizer and are linked to Wikidata.

TACRED: Unlike the previous two datasets, which were annotated using distant supervision, TACRED was *manually* annotated for 42 relation classes from the TAC KBP challenge (Surdeanu and Heng, 2014) (e.g., `per:schools_attended` and `org: members`) plus `no_relation`. The dataset contains 106,264 RE examples, which were annotated over textual data from both newswire sources and the corpus employed in the annual TAC Knowledge Base Population (TAC KBP) challenges (Zhang et al., 2017). These examples are generated by merging human annotations obtained from the TAC KBP challenges and crowdsourcing.

It is important to note that these datasets capture distinct phenomena that are important for RE:

(1) NYT29 and WIKIDATA were annotated using distant supervision, whereas TACRED was manually annotated. It is known that distant supervision introduces label noise (Riedel et al., 2010). This is particularly important for the negative class, i.e., in the case of distant supervision, negative labels can be false negatives. That is, they should not be interpreted as "no known relation label applies" but rather as "we have no information about this entity pair in the knowledge base." This impacts the sampling procedure discussed later in this section.

(2) NYT29 allows multiple relations to exist between the same two entities in the same sentence. For example, in the sentence "Mr. Mashal, speaking in Damascus, Syria, said …" and the entity pair "Damascus" and "Syria" is annotated with two relations: `administrative_divisions` and `capital`. Because of this, multi-label RE classifiers may have an advantage on NYT29.

(3) WIKIDATA allows for overlapping entities. For example, in the sentence "…featuring Lon Chaney and Andrew Lloyd Webber's 1986 musical ." and the entity pair "1986" and "1986 musical" is annotated with `first performance`. This is likely to confuse methods that rely on entity markers (Zhou and Chen, 2022).

3.2. Linguistic Annotations

Since some of these datasets were not accompanied by linguistic annotations, we processed the texts in house to guarantee that the same linguistic information is available for all three datasets. For all linguistic annotations, we used the `processors` library.[2] This library uses LSTM-CRFs (Lample et al., 2016) for case restoration, part-of-speech (POS) tagging, named entity recognition (NER), and the method of Vacareanu et al. (2020) for dependency parsing.

3.2.1. NYT29

In the original NYT29 dataset, the texts in the three partitions (train, dev, test) were initially presented in lowercase, which led to certain inaccuracies during linguistic annotation. To solve this problem, we first restored case using the LSTM-CRF in the `processors` library. On a small sample, we observed that this restoration is over 95% accurate.

We then tokenized the text and applied POS tagging, NER, and dependency parsing. However, to determine the subject and object type for each relation mentioned, we used the provided gold entity labels in the original dataset (see Table 1).

We observed that a small number of sentences in the NYT29 dataset were not parsed into a dependency tree by the `processors` parser (i.e., the parser produced several subtrees that covered different sentence fragments). The main cause of this error was long and complex sentences. However, the number of sentences with such errors was small: 0.1% of the training sentences, 0.07% in dev, and 0.1% in the test. For simplicity, we removed these sentences from train and dev, and, in order to not modify the test partition, we manually corrected the parse trees for the sentences in the test.

[2] https://github.com/clulab/processors

Sentence: "An arts center that the town of old Saybrook plans to open next year will be named after Katharine Hepburn."
Entity 1: "Katharine Hepburn" *Predicted label:* PERSON *Gold label:* PERSON
Entity 2: "old Saybrook" *Predicted label:* ORGANIZATION *Gold label:* LOCATION

Table 1: An example from NYT29 with gold and predicted entity labels. We used the gold entity labels for this dataset.

Dataset	Entity Labeling Scheme
NYT29	Gold labels
WIKIDATA	Predicted labels
TACRED	Predicted labels

Table 2: Labeling scheme for entities participating in relations in the three datasets considered.

3.2.2. WIKIDATA

For WIKIDATA, we used the same NLP library for tokenization, POS tagging, NER, and dependency parsing. Case restoration was not needed for the WIKIDATA sentences.

However, one important difference between NYT29 and WIKIDATA is that the labels for entities participating in relations in WIKIDATA are limited to just two: "Lexical" for named entities, and "Date" for dates. To increase the informativeness of entity labels, we adopted the labels predicted by the `processors` NER if they overlap with the span of the entity labels in WIKIDATA. If no predicted NE overlaps with a relation entity, we keep the default WIKIDATA entity label.

3.2.3. TACRED

In the TACRED dataset, essential NLP tasks, i.e., POS tagging, NER, and dependency parsing, were performed using Stanford CoreNLP (Manning et al., 2014) and included in the original dataset. To maintain compatibility with previous works, we keep the same linguistic annotations.

Importantly, TACRED and our version of WIKIDATA use labels predicted by a NER for the entities participating in a relation, whereas NYT29 uses gold labels. Table 2 summarizes this information.

3.3. Negative Class Label Standarization

The concept of negative relations refers to instances where the relation between two entities either does not fit into any predefined categories, or it may indicate that there is no relation between them at all. Note that negative labels are handled differently in the three datasets considered:

(1) NYT29 contains no annotations for the negative relation label. In this situation, we introduce negative examples using the supervised-to-few-shot transformation described in Section 3.4 and Algorithm 1.

(2) In contrast, TACRED and WIKIDATA explicitly annotate some negative relations between entity pairs that co-occur in the same sentence (TACRED uses the `no_relation` label, while WIKIDATA uses `P0`).

The above differences impact the few-shot version of these datasets (see § 3.4) and, thus, the performance of few-shot RE models. Lastly, we standardize the label for negative relations to `NOTA` across the three datasets.

To increase reproducibility, after all these preprocessing steps were applied, we formatted all three datasets using the same tabular format. The format is described in Appendix A. This is the same format the TACRED dataset used. We followed the exact format so that we could apply the transformation technique of converting the supervised dataset into the few-shot dataset described in (Sabo et al., 2021).

3.4. Supervised to Few-Shot Transformation

We transform the supervised NYT29, TACRED, and WIKIDATA datasets into FSRE datasets by applying a generalized form of the transformation method described in (Sabo et al., 2021). This process transforms a supervised dataset into a *realistic* FSRE dataset by following a series of constraints that are likely to occur in real-world applications:

(a) The test (or "target") relation classes are *different* from any of relations that might be available in a background dataset ("background relations");

(b) The number of training examples K for each target relation class is very small (typically 1 or 5);

(c) The distribution of relations is not uniform, i.e., some relations are rarer than others;

(d) Most candidate relation mentions do not correspond to a target relation;

(e) Many relation candidates seen in testing may not correspond also to a background relation. Thus, a traditional supervised RE classifier that trains on the background data is not applicable;

(f) Entities participating in relations may include named entities, as well as pronouns and common nouns.

Before we formalize the transformation process, we introduce some necessary notations:

$\boldsymbol{C}$ – A set of known relation classes in a dataset partition.

`NOTA` – The relation class `NOTA` (None-of-the-above) is assigned to entity pairs whose corresponding relation class is not in the applicable $\boldsymbol{C}$ set. Note that this is different from the `no_relation` label used in the supervised datasets. In the FSRE setting, `NOTA` includes both `no_relation` examples as well as all positive relation labels that are not used in the dataset partition at hand (Sabo et al., 2021).

$\boldsymbol{D}$ – A relation classification dataset such that $D: \{(x_i, c_i)\}_{i=1}^{n}$, where $\forall c_i \in C \cup \{\texttt{NOTA}\}$.

x_i – Represents the *i-th* instance in a RE dataset D such that $x_i = (e_1, e_2, s)_i$ where e_1 and e_2 represent a pair of entities in a sentence s, where the relation between this two entity is labeled c_i.

N**-Way** K**-Shot** – We follow the N-way K-shot setup for FSRE, as proposed by (Vinyals et al., 2016; Snell et al., 2017). In an N-way K-shot setup, a classifier aims to discriminate between N target relation classes using only a support set K examples of each. Typically, K is a very small number, e.g., 1 or 5.

Algorithm 1 describes the transformation process of a supervised RE dataset $\boldsymbol{D}$ containing relation labels $\boldsymbol{C}$ into a few-shot dataset $\boldsymbol{D}_{FS}$, $\boldsymbol{C}_{FS}$. The two key steps of the transformation algorithm are as follows. First, we split the original dataset into three partitions (train/dev/test) such that they are pairwise disjoint with respect to the positive relations they contain (steps 1 and 2). For example, if the train partition contains the relation `country of origin`, this relation is not allowed to appear in dev and test. Second, for each partition, we convert all relation labels that are assigned to another partition to `NOTA` (steps 3 and 4). Table 3 shows an example of the transformation process for WIKIDATA.

3.5. Episode Sampling

The small number of examples per class in FSRE (K) may introduce statistical instability in the results observed. To address this, episodic learning repeats the training/evaluation of a given method over a large number of episodes that sample different support sentences for the given classes. More formally, for a N-way K-shot setup an episode E consists of three items:

Algorithm 1 Transformation of a supervised RE dataset to few-shot RE using the N-way K-shot setup

Input: $\boldsymbol{D}, \boldsymbol{C}$
Output: $\boldsymbol{D}_{FS}, \boldsymbol{C}_{FS}$

Step 0: Replace `no_relation` with `NOTA` in $\boldsymbol{D}$; remove `no_relation` from $\boldsymbol{C}$, if present

Step 1: Split $\boldsymbol{D}$ in $\boldsymbol{D}_{train}$, $\boldsymbol{D}_{dev}$, and $\boldsymbol{D}_{test}$

Step 2: Choose a random split of $\boldsymbol{C}$ as $\boldsymbol{C}_{train}$, $\boldsymbol{C}_{dev}$, and $\boldsymbol{C}_{test}$ such that the following two conditions are true:

(a) $\boldsymbol{C}_{train}$, $\boldsymbol{C}_{dev}$, and $\boldsymbol{C}_{test}$ be pairwise disjoint
(b) $|\boldsymbol{C}_{train}|$, $|\boldsymbol{C}_{dev}|$, and $|\boldsymbol{C}_{test}| \geq N$ (for N-way K-shot)

Step 3:
for each $(x_i, c_i) \in \boldsymbol{D}_{train}$ **do**
  if $c_i \notin \boldsymbol{C}_{train}$ **then**
    $c_i = \text{NOTA}$
  else
    Retain the original label

Step 4: Repeat Step 3 for $\boldsymbol{D}_{dev}$ and $\boldsymbol{D}_{test}$ using their corresponding $\boldsymbol{C}_{dev}$, and $\boldsymbol{C}_{test}$ label sets

Step 5: $\boldsymbol{C}_{train} = \boldsymbol{C}_{train} \cup \{\text{NOTA}\}$
$\boldsymbol{C}_{dev} = \boldsymbol{C}_{dev} \cup \{\text{NOTA}\}$
$\boldsymbol{C}_{test} = \boldsymbol{C}_{test} \cup \{\text{NOTA}\}$

Step 6: $\boldsymbol{C}_{FS} = (\boldsymbol{C}_{train}, \boldsymbol{C}_{dev}, \boldsymbol{C}_{test})$
$\boldsymbol{D}_{FS} = (\boldsymbol{D}_{train}, \boldsymbol{D}_{dev}, \boldsymbol{D}_{test})$
return $\boldsymbol{C}_{FS}, \boldsymbol{D}_{FS}$

(a) N randomly chosen target relations:

$$\boldsymbol{C}_{target} = \{c_1, c_2, ..., c_N\} \text{ s.t. } c_{1..N} \notin \{\text{NOTA}\}$$

(b) A randomly chosen support set of size K for each of the N relations:

$$\boldsymbol{X}_{supt} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_i, ..., \boldsymbol{X}_{\boldsymbol{N}}\}$$

$$\boldsymbol{X}_i = \{(x_1, c_i), (x_2, c_i), ..., (x_j, c_i), .., (x_K, c_i)\}$$

(c) A randomly chosen labeled example as a query $Q = (x_q, c_q)$ such that $(x_q, c_q) \notin \boldsymbol{X}_{supt}$.

Given an episode $E = (\boldsymbol{C}_{target}, \boldsymbol{X}_{supt}, Q)$, the goal of a Few-Shot learning classifier is to create a decision function to choose a label from $C_{target} \cup \{\text{NOTA}\}$ for the given query Q.

We describe a general approach of N-Way K-Shot episode sampling procedure in the Algorithm 2, where $\boldsymbol{D}_E$ and $\boldsymbol{C}_E$ are input dataset and labels to sample from, and $\boldsymbol{E}_{test}$ is the set of returned episodes.

A similar episode sampling approach has been described in (Sabo et al., 2021). To create *train* episodes, D_{train} and C_{train} should be used as input. In the same way, *dev* and *test* episodes can be created using their respective data and relation sets.

Sentence 1: "Among the current participants, Iceland, Norway, and Switzerland are not members of the European Union." **Entity pair:** "Norway", "Switzerland" **Original label:** `no_relation` **Label after transformation:** `NOTA` **Reason:** `no_relation` in the supervised setting becomes `NOTA` for FSRE
Sentence 2: "Horror writer Stephen King once visited his friend, Peter Straub, whose house is in Crouch End." **Entity pair:** "Peter Straub", "Crouch End" **Original label:** `residence` **Label after transformation:** `residence` **Reason:** The sentence is in the dev set, and the relation label `residence` is part of $\boldsymbol{C}_{dev}$
Sentence 3: "Progeny is an American science fiction film released in 1999." **Entity pair:** "Progeny", "American" **Original label:** `country of origin` **Label after transformation:** `NOTA` **Reason:** The sentence is taken from the dev set, but the relation `residence` is part of $\boldsymbol{C}_{test}$

Table 3: Example data points before and after the transformation process in Algorithm 1.

Algorithm 2 Episode sampling for a N-way K-shot FSRE

Input: $\boldsymbol{D}_E, \boldsymbol{C}_E, episodeSize$
Output: $\boldsymbol{E}_{test}$
$\boldsymbol{E}_{test} = \{\}$
$\boldsymbol{C}'_E = \boldsymbol{C}_E - \{\text{NOTA}\}$
for $e = 0$ to $episodeSize$ **do**
  $\boldsymbol{C}_{target} = RandomSample(\boldsymbol{C}'_E, N)$
  $\boldsymbol{X}_{supt} = [\,]$
  for $i = 0$ to $|C_{target}|$ **do**
    $r = C_{target}[i]$
    $\boldsymbol{X}_i = RandomSample(\boldsymbol{D}_E, K, r)\}$
    $\boldsymbol{X}_{supt}[i] = \boldsymbol{X}_i$
  $\boldsymbol{D}'_E = \{\boldsymbol{D}_E : \boldsymbol{D}_E \notin \boldsymbol{X}_{supt}\}$
  $Q = RandomSample(\boldsymbol{D}'_E, 1)$
  $\boldsymbol{C}'_{target} = \boldsymbol{C}_{target} \cup \{\text{NOTA}\}$
  $\boldsymbol{E}_{test} = \boldsymbol{E}_{test} \cup \{(\boldsymbol{C}'_{target}, X_{supt}, Q)\}$
return $\boldsymbol{E}_{test}$

4. Dataset Statistics

Table 4 summarizes key statistics for the three supervised datasets that serve as the starting point for FSRE. We chose three datasets with a significant variation in the number of relations. Table 4 shows that TACRED has 42 relations, NYT29 has 29 relations, and WIKIDATA has 352 relations, which is much larger. Additionally, when we look at the NOTA instances, these three datasets differ enormously. For instance, in the supervised NYT29, there are no NOTA instances. In supervised TACRED, the number of NOTA instances is higher than the number of NOTA instances in the supervised WIKIDATA. Table 5 summarizes how the number of relation instances and NOTA instances in three resulting FS datasets have been changed from supervised datasets. In Appendix B, we present further statistics and analysis demonstrating that our FSRE meta-dataset meets all the requirements of a realistic few-shot relation extraction dataset.

	TACRED	NYT29	WIKIDATA
Train Size	68,124	78,885	775,919
Dev Size	22,631	5859	251,802
Test Size	15,509	8759	739,408
Relation Class	42	29	352
Relation Instances	21,773	93,503	1,299,085
NOTA Instances	84,491	0	468,044

Table 4: Statistics of the supervised TACRED, NYT29, and WIKIDATA datasets. The first three rows report the number of sentences per partition.

	TACRED	NYT29	WIKIDATA
Relation Instances	9,600	58,841	513,891
NOTA Instances	96,664	34,662	1,253,238

Table 5: Statistics of the Few-Shot TACRED, NYT29, and WIKIDATA datasets.

5. Experimental Results

5.1. Experimental Setup

We applied the transformation techniques outlined in § 3.4 and § 3.5 on all datasets described in the previous section to produce their FSRE variants.[3] We tested on all datasets in 5-way 1-shot and 5-way 5-shot scenarios. In both cases, we repeat the procedure with 5 different random seeds.

5.2. Models

We evaluated the following baselines and models:

[3]To enable comparison with previous work, for TACRED we kept the transformation introduced in (Sabo et al., 2021).

Unsupervised Baseline – This baseline model uses *solely* the entity types in both the query sentence and the support sentences for classification during inference (Vacareanu et al., 2022b). If there are support sentences with the same entity types as the query sentence, the model randomly chooses one and predicts its relation. In other cases, the baseline predicts NOTA.

Sentence-Pair – We implement a baseline similar to (Gao et al., 2019), which operates as follows: We pair each query sentence to each support sentence and feed the concatenated text to a sentence transformer (Reimers and Gurevych, 2019) to obtain a single score that quantifies the degree to which both sentences convey the same underlying relation. During training, we fine-tune the model to maximize the score between sentences with the same relation and minimize the score between sentences with different relation (or NOTA). During inference, we predict the relation associated with the highest score, provided it is above a threshold tuned on the development partition. Otherwise, we predict NOTA. We use a pre-trained model and show results with and without fine-tuning.[4]

MNAV – Multiple NOTA Vectors (MNAV) is an extended version of the NAV method, which computes a score between the query vector, each support sentence vector, and, additionally, a learned vector for the NOTA class (Sabo et al., 2021). Instead of just one vector for NOTA, MNAV uses multiple vectors to account for the fact that NOTA is a "catch all" for all other relations. The number of NOTA vectors is tuned on the development set. In the classification process, the model selects the nearest vector to the query representation to establish the predicted relation label.

OdinSynth – OdinSynth is a transformer-based rule synthesis model that generates rules from the provided support sentences and then applies these rules to the query sentence (Vacareanu et al., 2022b). If none of the rules match, the model predicts NOTA. If there exists a match with one or more rules, the model predicts the relation through majority voting.

Hard-Matching Rules – Represent lexico-syntactic rules created over the shortest path connecting the two entities.

Soft-Matching Rules – This is a neuro-symbolic model (Vacareanu et al., 2024) that aims to increase the recall of rules by leveraging the high expressivity of neural networks. The method first

[4]`cross-encoder/ms-marco-MiniLM-L-6-v2`

Model	5-way 1-shot			5-way 5-shot		
	P	R	F1	P	R	F1
Unsupervised Baseline	5.70 $\pm$ 0.10	91.02 $\pm$ 0.65	10.73 $\pm$ 0.18	5.65 $\pm$ 0.11	95.56 $\pm$ 0.70	10.67 $\pm$ 0.20
Sentence-Pair (not fine-tuned)	3.9 $\pm$ 0.21	5.21 $\pm$ 0.31	4.45 $\pm$ 0.24	2.76 $\pm$ 0.16	8.79 $\pm$ 0.58	4.2 $\pm$ 0.25
Sentence-Pair (fine-tuned)	6.89 $\pm$ 0.33	28.56 $\pm$ 1.67	11.10 $\pm$ 0.55	14.94 $\pm$ 0.26	24.03 $\pm$ 0.32	18.42 $\pm$ 0.16
MNAV	15.11 $\pm$ 0.46	8.47 $\pm$ 0.31	10.85 $\pm$ 0.29	24.48 $\pm$ 1.02	32.00 $\pm$ 1.07	27.73 $\pm$ 0.94
OdinSynth	23.48 $\pm$ 1.46	11.46 $\pm$ 1.02	15.40 $\pm$ 1.21	29.77 $\pm$ 0.83	20.34 $\pm$ 0.53	24.16 $\pm$ 0.44
Hard-matching Rules	51.35 $\pm$ 6.53	2.94 $\pm$ 0.48	5.56 $\pm$ 0.90	45.94 $\pm$ 5.31	10.81 $\pm$ 1.23	17.50 $\pm$ 1.98
Soft-matching Rules	33.46 $\pm$ 1.47	19.69 $\pm$ 1.14	**24.78 $\pm$ 1.22**	51.66 $\pm$ 1.85	26.02 $\pm$ 1.29	**34.59 $\pm$ 1.24**

Table 6: The results for the 5-way 1-shot and 5-way 5-shot settings on the test partition of the FS TACRED dataset.

Model	5-way 1-shot			5-way 5-shot		
	P	R	F1	P	R	F1
Unsupervised Baseline	11.60 $\pm$ 0.18	40.34 $\pm$ 0.54	18.03 $\pm$ 0.26	11.70 $\pm$ 0.25	40.65 $\pm$ 0.45	18.17 $\pm$ 0.34
Sentence-Pair (not fine-tuned)	10.61 $\pm$ 0.32	12.39 $\pm$ 0.41	11.43 $\pm$ 0.35	15.81 $\pm$ 0.94	5.41 $\pm$ 0.25	8.06 $\pm$ 0.39
Sentence-Pair (fine-tuned)	38.09 $\pm$ 2.42	7.4 $\pm$ 0.42	12.4 $\pm$ 0.71	36.48 $\pm$ 1.37	16.02 $\pm$ 0.41	**22.26 $\pm$ 0.62**
MNAV	25.08 $\pm$ 0.73	34.37 $\pm$ 0.87	**29.00 $\pm$ 0.80**	33.24 $\pm$ 1.06	15.47 $\pm$ 0.38	21.12 $\pm$ 0.55
OdinSynth	30.07 $\pm$ 0.93	9.42 $\pm$ 0.31	14.34 $\pm$ 0.46	21.61 $\pm$ 0.61	17.98 $\pm$ 0.45	19.63 $\pm$ 0.51
Hard-matching Rules	77.47 $\pm$ 1.53	1.53 $\pm$ 0.13	3.01 $\pm$ 0.25	80.49 $\pm$ 1.73	3.40 $\pm$ 0.12	6.52 $\pm$ 0.23
Soft-matching Rules	20.80 $\pm$ 0.38	12.27 $\pm$ 0.39	15.44 $\pm$ 0.40	24.50 $\pm$ 0.83	16.67 $\pm$ 0.49	19.84 $\pm$ 0.59

Table 7: The results for the 5-way 1-shot and 5-way 5-shot settings on the test partition of the FS NYT dataset.

Model	5-way 1-shot			5-way 5-shot		
	P	R	F1	P	R	F1
Unsupervised Baseline	2.52 $\pm$ 0.16	29.99 $\pm$ 1.42	4.64 $\pm$ 0.28	2.28 $\pm$ 0.13	54.35 $\pm$ 1.03	4.38 $\pm$ 0.24
Sentence-Pair (not fine-tuned)	6.4 $\pm$ 1.51	2.55 $\pm$ 0.66	3.65 $\pm$ 0.92	2.68 $\pm$ 0.56	8.67 $\pm$ 1.57	4.09 $\pm$ 0.82
Sentence-Pair (fine-tuned)	6.65 $\pm$ 0.78	7.99 $\pm$ 0.93	7.26 $\pm$ 0.85	5.76 $\pm$ 0.87	8.74 $\pm$ 0.95	6.94 $\pm$ 0.93
MNAV	17.49 $\pm$ 1.45	6.76 $\pm$ 1.21	**9.74 $\pm$ 1.47**	15.27 $\pm$ 0.98	28.26 $\pm$ 0.96	**19.83 $\pm$ 1.06**
OdinSynth	12.99 $\pm$ 1.67	6.15 $\pm$ 0.58	8.34 $\pm$ 0.85	10.09 $\pm$ 1.31	19.18 $\pm$ 1.57	13.21 $\pm$ 1.46
Hard-matching Rules	6.38 $\pm$ 3.24	0.38 $\pm$ 0.20	0.72 $\pm$ 0.37	5.15 $\pm$ 1.83	1.13 $\pm$ 0.37	1.85 $\pm$ 0.61
Soft-matching Rules	35.88 $\pm$ 10.01	2.73 $\pm$ 0.86	5.06 $\pm$ 1.58	17.58 $\pm$ 3.28	9.71 $\pm$ 2.15	12.50 $\pm$ 2.59

Table 8: The results for the 5-way 1-shot and 5-way 5-shot settings on the test partition of the FS WIKIDATA dataset.

attempts to match a rule the traditional way (see *Hard-Matching Rules*). If the match fails, it then falls back to the neural component, which will predict a matching score $s \in [0, 1]$. The training of the neural component utilizes (rule, sentence) tuples along with a contrastive loss function. The objective is to maximize the similarity between rules and sentences with the same relation while minimizing it for those with different relations.

In addition to a comprehensive assessment of the six recent few-shot relation extraction models mentioned above, we also evaluated a Zero-Shot Large Language Model (LLM) baseline on our FSRE meta-dataset. The details of this baseline and the experimental result are provided in Appendix C.

5.3. Results Analysis

Table 6, 7, 8 represent the result of different models on our resulting FSRE datasets. We draw the following conclusions:

First, no single method emerges as the clear top performer across all scenarios. For instance, *Soft-matching Rules* achieves the highest performance on Few-Shot TACRED (Table 6), *MNAV* excels on Few-Shot WIKIDATA (Table 8), and in the case of Few-Shot NYT, *MNAV* performs best for 1-shot, while *Sentence-Pair* leads for 5-shot (Table 7). The latter result is surprising, given the simplicity of this baseline.

Second, the performance varies drastically between the datasets for both 1-shot and 5-shot scenarios. For instance, in Few-Shot WIKIDATA 5-way 1-shot, the top-performing method achieves an F1 score of 9.74, whereas in Few-Shot TACRED 5-way 1-shot, the best method reaches an F1 score of 24.78. This underscores the importance of employing multiple evaluation datasets to gain a realistic assessment of a model's performance. Further, the overall performance across all datasets is low, which indicates a substantial need for future research in this domain.

In our evaluation of the six models, FS WIKIDATA exhibited comparatively lower performance across all datasets. To understand the underlying reasons, we conducted a qualitative error analysis on FS WIKIDATA, the details of which are provided in Appendix D.

6. Conclusion

In this paper, we presented a meta dataset for few-shot relation extraction (FSRE), which comprises three FSRE datasets: two were derived from established supervised relation extraction datasets, while one is an existing FSRE dataset. All datasets were intricately derived to replicate real-world scenarios, ensuring a strong alignment with real-world contexts. Then, we assessed six relation extraction methods on this meta dataset and found that no single model consistently performs well across all scenarios. This suggests the need for future research in this domain.

As future work, we plan to leverage the resulting dataset to develop methods that demonstrate consistent and robust performance.

7. Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus and Automating Scientific Knowledge Extraction and Modeling (ASKEM) programs. Mihai Surdeanu declares a financial interest in `lum.ai`. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

8. Bibliographical References

Eneko Agirre. 2022. Few-shot information extraction is here: Pre-train, prompt and entail. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Elisa Bassignana and Barbara Plank. 2022. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lihu Chen, Simon Razniewski, and Gerhard Weikum. 2023. Knowledge base completion for long-tail entities. *arXiv preprint arXiv:2306.17472*.

Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, and Dan Roth. 2022. New frontiers of information extraction. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*.

Amir Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation extraction as two-way span-prediction. *ArXiv*, abs/2010.04829.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Conference on Empirical Methods in Natural Language Processing*.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

S. Gupta and Christopher D. Manning. 2014. Improved pattern learning for bootstrapped entity extraction. In *CoNLL*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Y. Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Hany Hassan, Ahmed Hassan Awadallah, and Ossama Emam. 2006. Unsupervised information extraction approach using graph mutual reinforcement. In *EMNLP*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yu Su Kai Zhang, Bernal Jiménez Gutiérrez. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*.

Zornitsa Kozareva, Ellen Riloff, and Eduard H. Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *ACL*.

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

K. Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. *ArXiv*, abs/2205.09837.

Christopher D. Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41:701–707.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Neural Information Processing Systems*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8528–8535.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *VS@HLT-NAACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *ArXiv*, abs/1802.05365.

Mahdi Rahimi and Mihai Surdeanu. 2023. Improving zero-shot relation classification via automatically-acquired entailment templates. In *Workshop on Representation Learning for NLP*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Ellen Riloff. 1993. Automatically constructing a dictionary for information extraction tasks. In *AAAI*.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *AAAI/IAAI, Vol. 2*.

Ellen Riloff and R. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *EMNLP*.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. *ArXiv*, abs/2109.03659.

Evan Sandhaus. 2008. The new york times annotated corpus. Web Download. LDC2008T19.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789.

Mihai Surdeanu and Ji Heng. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proceedings of the TAC-KBP 2014 Workshop*.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference (NAACL-2010)*, Los Angeles, CA.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7072–7079.

Zheng Tang and Mihai Surdeanu. 2021. Interpretability rules: Jointly bootstrapping a neural relation extractorwith an explanation decoder. *Proceedings of the First Workshop on Trustworthy Natural Language Processing*.

Zheng Tang and Mihai Surdeanu. 2023. Bootstrapping neural relation and explanation classifiers. In *Annual Meeting of the Association for Computational Linguistics*.

Robert Vacareanu, Fahmida Alam, Md Asiful Islam, Haris Riaz, and Mihai Surdeanu. 2024. Best of both worlds: A pliable and generalizable neuro-symbolic approach for relation classification. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Robert Vacareanu, George C. G. Barbosa, Marco A. Valenzuela-Escarcega, and Mihai Surdeanu. 2020. Parsing as tagging. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*.

Robert Vacareanu, Dane Bell, and Mihai Surdeanu. 2022a. Patternrank: Jointly ranking patterns and extractions for relation extraction using graph-based algorithms. In *PANDL*.

Robert Vacareanu, Marco A. Valenzuela-Escárcega, George Caique Gouveia Barbosa, Rebecca Sharp, Gustave Hahn-Powell, and Mihai Surdeanu. 2022b. From examples to rules: Neural guided rule synthesis for information extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6180–6189, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *International Conference on Computational Linguistics*.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *ArXiv*, abs/1508.01006.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

Wenxuan Zhou, Hongtao Lin, Bill Yuchen Lin, Ziqi Wang, Junyi Du, Leonardo Neves, and Xiang Ren. 2020. Nero: A neural rule grounding framework for label-efficient relation extraction. *The Web Conference '20 in arXiv: Computation and Language*.

9. Language Resource References

Tianyu Gao and Xu Han and Hao Zhu and Zhiyuan Liu and Peng Li and Maosong Sun and Jie Zhou. 2019. *FewRel 2.0: Towards More Challenging Few-Shot Relation Classification*. [link].

Xu Han and Hao Zhu and Pengfei Yu and Ziyun Wang and Y. Yao and Zhiyuan Liu and Maosong Sun. 2018. *FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation*. [link].

Evan Sandhaus. 2008. *The New York Times Annotated Corpus*. Linguistic Data Consortium.

Sorokin, Daniil and Gurevych, Iryna. 2017. *Context-aware representations for knowledge base relation extraction*.

Surdeanu, Mihai and Heng, Ji. 2014. *Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation*.

Zhang, Yuhao and Zhong, Victor and Chen, Danqi and Angeli, Gabor and Manning, Christopher D. 2017. *Position-aware attention and supervised data improve slot filling*.

Appendix A

Table 9 summarizes the tabular format used to represent the three supervised datasets that are the starting point of the few-shot datasets generated in this work.

Field	Description
id	Incremental unique ID for each example or sentence
docid	For dev set docid = "dev", for test set docid= "test", and for train set docid = "train"
relation	This field denotes the relation labels between the given entities
token	An instance of a sequence or word in the sentence
subj_start	Start index of the subject in a sentence
subj_end	End index of the subject in a sentence
obj_start	Start index of the object in a sentence
obj_end	End index of the object in a sentence
subj_type	Subject type (e.g., person name) in a sentence
obj_type	Object type (e.g., person name) in a sentence
stanford_pos	POS tag of the current token
stanford_ner	Named entity label of the current token
stanford_head	1-based index of the dependency head of the current token
stanford_deprel	dependency relation of the current token to its head token

Table 9: Descriptions of the columns in the tabular format used to encode the three supervised datasets used in this work. Note that the "stanford" prefix for the last three columns is maintained for compatibility with the TACRED format; in NYT29 and WIKIDATA, this information is generated using the `processors` library instead.

Appendix B

In section 3.4, we outlined six characteristics essential for a realistic Few-Shot dataset. Our FSRE

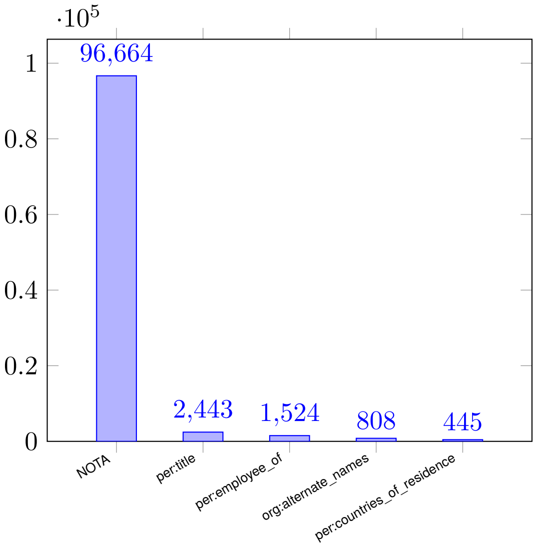

Figure 1: Few-Shot TACRED top five relation distribution

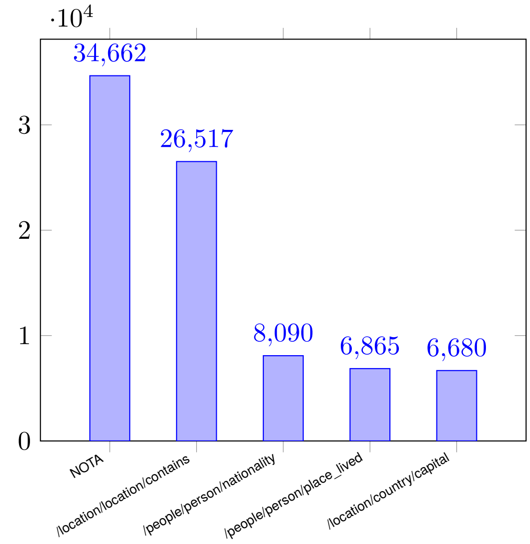

Figure 2: Few-Shot NYT29 top five relation distribution

meta dataset fulfills all these six criteria. For instance, we split the original dataset into three partitions (train/dev/test) such that they are pairwise disjoint with respect to the positive relations they contain (as outlined in Steps 1 and 2 of Algorithm 1). This guarantees that the test relation classes in our dataset are distinct from any relations that might be present in a background dataset, thereby fulfilling constraint (a) of the realistic Few-Shot assumption. Moreover, we follow the 5-way 1-shot and 5-way 5-shot setup for episode sampling, which ensures that the number of training examples for each target relation class is very small (1 or 5), thus satisfying constraint (b). Figure 1, 2, 3 illustrate the non-uniformity of the relation classes and the predominance of `NOTA` class, indicative of satisfying realistic constraints (c), (d), and (e). Figure 4, 5, 6 indicate the presence of a variety of POS tags, with a notable percentage of proper nouns, common nouns, and pronouns, reflecting the diversity and realism of entity distributions, thus satisfying constraint (f).

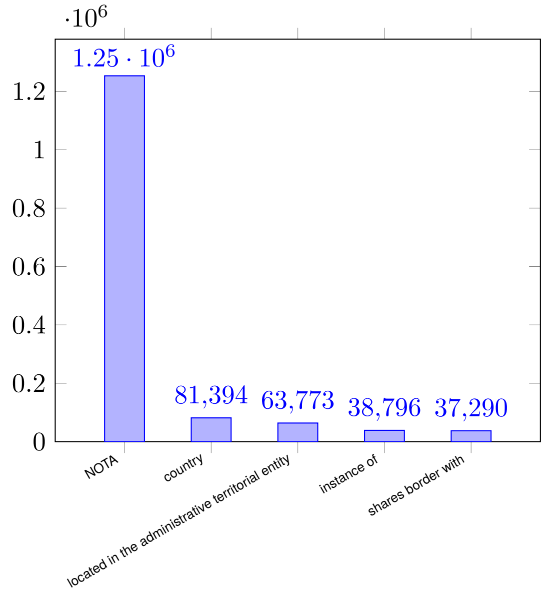

Figure 3: Few-Shot WIKIDATA top five relation distribution

Appendix C

Zero-Shot LLM Baseline

We evaluated the Zero-Shot relation classification performance of the Large Language Model (LLM) using *GPT-4*. The experiment was conducted on a test set containing ten episodes, with each episode containing three test sentences. For each sentence, we prompted *GPT-4* to identify a relation for a given entity pair using the prompting technique described by Kai Zhang (2023). The prompt includes the label verbalization technique to articulate the relations. We conducted the experiment in both 5-way 1-shot and 5-way 5-shot configurations, where *GPT-4* was tasked with classifying the relation for the given entity pair into one of the five target relations or indicating 'None of the Above' (`NOTA`) if none is applicable. Figure 7 illustrates an example of the prompt.

The results of the experiment are presented in Tables 10, 11 and 12. In the tables, we included the performance scores of other models on the same test set to facilitate easier comparison. The results show that zero-shot LLM achieves low precision and high recall in FS TACRED (see Table 10) and FS NYT29 (see Table 11). The low precision is attributed to a high false positive rate, where

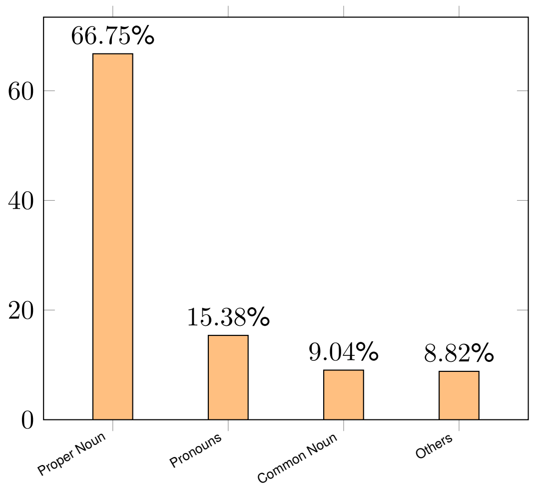

Figure 4: Few-Shot TACRED Entity POS tag distributions

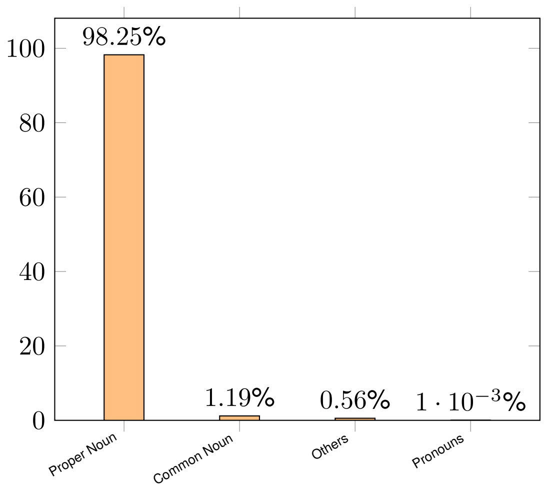

Figure 5: Few-Shot NYT29 Entity POS tag distributions

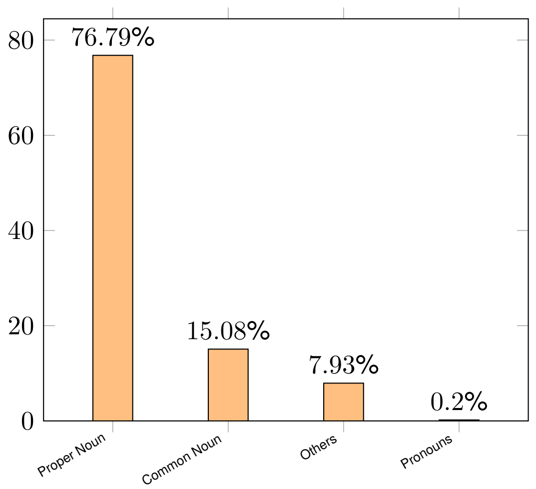

Figure 6: Few-Shot WIKIDATA Entity POS tag distributions

```
Given a sentence and two entities
within the sentence, classify the
relation between the two entities
based on the provided sentence. All
possible relations are listed below:

-org:top_members/employees: Entity
 1 has the high level member Entity 2

-per:schools_attended: Entity 1
 studied in Entity 2

-org:founded_by: Entity 1 was founded
 by Entity 2

-per:origin: Entity 1 has the
 nationality Entity 2

-per:date_of_birth: Entity 1 has
 birthday on Entity 2

-NOTA: None of the above


 Sentence: "In an atmosphere of
 conflict and misunderstanding, the
 travel and tourism industry can be
 an incredibly powerful force for
 conciliation," said PATA president
 and chief executive officer Peter de
 Jong.

 Entity 1: PATA
 Entity 2: Peter de Jong
```

Figure 7: An example of prompt for Zero-Shot LLM baseline.

GPT-4 often chose a positive relation from the target set instead of selecting `NOTA` when the correct relation was not among the target relations. However, when the correct relation is included in the target set, *GPT-4* tends to identify it correctly, resulting in a high true positive rate. Although GPT-4 generally performs better than the other models, it is not always the best in every scenario. For example, in the FS NYT29 5-way 1-shot configuration, the *MNAV* model outperforms *GPT-4*, and in the FS WIKIDATA 5-way 5-shot setup, the *Unsupervised Baseline* model performs better than GPT-4. This reinforces the conclusion drawn in section 5.3 that no single model consistently stands out as the best performer across all scenarios, underscoring the significant need for continued research in this field.

Since a small test set was utilized in this experiment, further research is necessary to gain a deeper understanding of the Zero-Shot relation classification capabilities of Large Language Models.

Appendix D

Qualitative Error Analysis

In the few-shot relation extraction (FSRE) setting, the performance of all six models we evaluated was comparably low when evaluated on WIKI-

Model	5-way 1-shot			5-way 5-shot		
	P	R	F1	P	R	F1
Unsupervised Baseline	8.33	33.33	13.33	11.76	66.67	20
MNAV	0	0	0	25	33.33	28.57
Hard-matching Rules	0	0	0	0	0	0
Soft-matching Rules	66.67	66.67	66.67	16.67	33.33	22.22
Zero-Shot LLM (GPT 4)	50	100	**67**	27	100	**43**

Table 10: The results for the 5-way 1-shot and 5-way 5-shot settings on a small test partition of the FS TACRED dataset.

Model	5-way 1-shot			5-way 5-shot		
	P	R	F1	P	R	F1
Unsupervised Baseline	18.18	50	26.67	12.5	37.50	18.75
MNAV	58.33	87.5	**69.99**	10.07	55.77	17.06
Hard-matching Rules	0	0	0	0	0	0
Soft-matching Rules	25	37.5	30	25	37.5	30
Zero-Shot LLM (GPT 4)	26.08	75	38.71	21.74	62.5	**32.26**

Table 11: The results for the 5-way 1-shot and 5-way 5-shot settings on a small test partition of the FS NYT29 dataset.

Model	5-way 1-shot			5-way 5-shot		
	P	R	F1	P	R	F1
Unsupervised Baseline	10	10	10	50	58.33	**53.85**
MNAV	0	0	0	66.67	16.67	26.67
Hard-matching Rules	100	0	0	100	0	0
Soft-matching Rules	33.33	10	15.38	33.33	10	15.38
Zero-Shot LLM (GPT 4)	55.55	50	**52.63**	60	46.15	52.17

Table 12: The results for the 5-way 1-shot and 5-way 5-shot settings on a small test partition of the FS WIKIDATA.

DATA. This can be primarily attributed to the high prevalence of long-tail entities in WIKIDATA. In (Chen et al., 2023), it is reported that approximately half of the entities in WIKIDATA fall into the long-tail category. The challenges stemming from this prevalence of long-tail entities contribute significantly to the observed performance degradation. Firstly, the data scarcity inherent in long-tail entities exacerbates the already challenging few-shot learning scenario, where models are expected to generalize from limited examples. With fewer instances available for these long-tail entities, models struggle to capture the diverse range of relation patterns and semantic nuances associated with them. Additionally, the lack of contextual cues and varied semantic contexts surrounding these entities further compounds the difficulty of accurate relation extraction. As a result, the efficacy of models in the FSRE setting is hampered by the combination of data scarcity and the intricate nature of relations involving long-tail entities in WIKIDATA.