A Study on Building Efficient Zero-Shot Relation Extraction Models

Hugo Thomas Caio Corro Guillaume Gravier Pascale Sébillot
Univ Rennes, INSA Rennes, CNRS, Inria, IRISA – UMR 6074, F-35042 Rennes, France
{firstname.lastname}@irisa.fr

Abstract

Zero-shot relation extraction aims to identify relations between entity mentions using textual descriptions of novel types (i.e., previously unseen) instead of labeled training examples. Previous works often rely on unrealistic assumptions: (1) pairs of mentions are often encoded directly in the input, which prevents offline pre-computation for large scale document database querying; (2) no rejection mechanism is introduced, biasing the evaluation when using these models in a retrieval scenario where some (and often most) inputs are irrelevant and must be ignored. In this work, we study the robustness of existing zero-shot relation extraction models when adapting them to a realistic extraction scenario. To this end, we introduce a typology of existing models, and propose several strategies to build single pass models and models with a rejection mechanism. We adapt several state-of-the-art tools, and compare them in this challenging setting, showing that no existing work is really robust to realistic assumptions, but overall AlignRE (Li et al., 2024b) performs best along all criteria.

Keywords: relation extraction, zero-shot learning, efficient models, rejection mechanism

1. Introduction

Relation extraction (RE) is a fundamental natural language processing task that aims to identify which relation links two entity mentions, if any—see Figure 1 (left). RE is central in many fields including biomedical research (Habibi et al., 2017), finance (Hamad et al., 2024) and legal analysis (Zhong et al., 2020), mainly as an intermediate step for solving downstream tasks, including in information retrieval (Yang et al., 2022), question answering (Chen et al., 2019), document retrieval for clinical decision process (Agosti et al., 2018), etc.

Although several surveys on RE have been published (Hang et al., 2025; Zhao et al., 2024; Deng et al., 2024; Bassignana and Plank, 2022, *inter alia*), we argue that there is a need to better understand existing models in the context of specific use cases. Indeed, users may face challenging settings that encompass constraints that have been little studied in the literature. In this work, we focus on zero-shot RE, which we cast as a relation mining problem where we assume the targeted text dataset is large but known in advance, e.g., a company's document archive. We illustrate this use case with the following scenario.

> **Use case scenario.** A journalist wants to search for specific facts in a collection of raw news archives. He starts by describing relation types of interest: "`country in which this person rigged elections`"; "`person who illegally financed elections in this country`"; etc. Then, he uses a RE model to retrieve all occurrences of these relations in the archives.

This setting is particularly challenging. First, the targeted relation types are not known in advance, which we call *on-the-fly* zero-shot RE, in order to insist that each novel query on the data may request for previously unknown types. Second, to ensure reasonable processing times, models must allow to pre-compute and store input text representations. In this work, we refer to this property as *offline encoding*.

Definition 1 (On-the-fly zero-shot classification)**.** *On-the-fly zero-shot classification refers to the problem of multiclass classification for which the set of output classes is (1) unseen during training and (2) instance dependent, that this the set of output classes is specified at the same time as the instance to classify; in other words the set of output classes is part of the input.*

Definition 2 (Offline encoding)**.** *Offline encoding refers to the setting where representations of instances (e.g., contextual embeddings of tokens for textual inputs) are pre-computed and stored in advance.*

These two requirements constrain the neural architecture to be based on *late interaction* (Khattab and Zaharia, 2020), that is the encoding step (e.g., using Bert or related models) must compute representations of the input utterance and the relation types separately, and only combine them when computing output logits, see Figure 2. In addition to late interaction, offline encoding means that entity mention candidates cannot be identified in the input during the utterance encoding step.

Last, when searching for relations of a given type in a large collection, most inputs should be rejected as they are irrelevant, and therefore the standard multi-class classification setting is inadequate.[1]

[1]For example, the second instance in Figure 1 is

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6056–6067
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

operating system

(1) Cliqz supports the macOS operating system.
head: Cliqz — tail: macOS

(2) Cliqz does not support ArchLinux.
head: Cliqz — tail: ArchLinux.

Side-information on the relation type:

- **Name:** "operating system"
- **Aliases:** "OS, "supported OS", ...
- **Description:** "operating system (OS) on which a software works or the OS installed on hardware"

Figure 1: In relation extraction, the input is an utterance with two identified mentions, a head mention and a tail mention. We assume that the only targeted relation type is `operating system`. **(left)** Two input examples. In (1), the model must predict that there is a relation of this type between the two mentions. In (2), the model should reject the input, as the candidate input relation does not correspond to any type in the targeted ones. **(right)** Example of side-information used for zero-shot relation extraction.

This requires to augment the model with a *rejection mechanism* (Hendrycks and Gimpel, 2017; Barandas et al., 2022; Hendrickx et al., 2024).

> **Research question.** Can we easily identify the best model for *on-the-fly* and *offline encoding* RE in large text collections, and how robust are these models when enhanced with the necessary *rejection mechanism* for this use-case?

To answer this question, we propose a novel typology that shows that no recent off-the-shelf model is tailored to our use case. Our classification shows that most approaches in the literature encode the targeted relation directly in the input by explicitly augmenting the input utterance to identify the head and tail mentions candidate of a relation: for instance, utterance (1) in Figure 1 can have its entities targeted using special markup: "`<E1>Cliqz</E1> supports the <E2>macOS</E2> operating system.`". This means that these models do not enable offline encoding, even when they are based on a late-interaction architecture.[2] Moreover, to the best of our knowledge, our typology shows that no recent zero-shot RE based on a late-interaction architecture implements a "native" rejection mechanism.

Using our typology, we identify three state-of-the-art models whose source codes are publicly available: EMMA (Li et al., 2024a), REMATCHING (Zhao et al., 2023) and ALIGNRE (Li et al., 2024b). We precisely describe the differences between these models, and explain how to adapt them to our use case, i.e., to allow offline encoding and on-the-fly zero-shot RE. As all these models lack a rejection mechanism, we present three different options based on previous works. Finally, we evaluate these updated models with and without rejection mechanisms, on two publicly available datasets, FEWREL (Han et al., 2018) and WIKIZSL (Chen and Li, 2021).

Our experiments show that ALIGNRE, when adapted for offline encoding and a rejection mechanism, performs best among all considered models.

Our contributions can be summarized as follows: (1) we build a typology of zero-shot RE models; (2) we compare the main differences in the architecture of three state-of-the art models after adjustment for offline encoding; (3) we describe three rejection mechanisms that can be implemented in any zero-shot RE model; (4) we evaluate these models in comparable evaluation settings. Code to reproduce experiments is publicly available.[3]

2. Related Works

In this section, we review previous publications following similar goals as ours, giving a clear view over the current state of models and identifying design choices that impact downstream results.

2.1. Related Typologies

Building a typology, i.e., a set of targeted criteria, enables systematic model comparison and selection for specific scenarios.

Zhao et al. (2024) focus on *supervised* RE, whereas we focus on zero-shot learning. Closer to our setting, Hang et al. (2025) introduced a taxonomy of manual and automatic prompts for few-shot RE. They analyze template construction and model fine-tuning strategies, as well as their pros and cons (e.g., annotation costs, prediction time,

rejected because the relationship between the two entity mentions does not correspond with to of the target relation types.

[2]RE focuses on classifying a relation candidate identified by two mentions. In the zero-shot learning scenario, the targeted types may concern mention types that are unknown during the pre-computation stage. Note however that, as in previous zero-shot RE works, for evaluation we assume the gold mentions are given.

[3]https://gitlab.inria.fr/huthomas/zsre-models-adaptation

etc). We focus on other criteria, namely processing efficiency and rejection methods.

Beyond RE, Deng et al. (2024) and Pai et al. (2024) studied low-resource and open information extraction, respectively, but do not consider zero-shot RE. We take inspiration from these studies, differentiating ourselves by the goal of our typology, that is finding adapted or adaptable models to large scale and on-the-fly relation extraction thanks to offline encoding.

2.2. Datasets and Evaluation

Bassignana and Plank (2022) proposed a broad overview on RE datasets and evaluation protocols. Importantly, they show that annotation is not consistent across datasets. However, their analysis does not mention the presence or absence of side-information about relation types (like textual description, aliases, etc.) that are mandatory for our zero-shot setting.

Han et al. (2018) introduced the FewRel dataset for few-shot RE. Relation types and their instances are based on Wikidata. Chen and Li (2021) proposed to use the same dataset for zero-shot RE by simply changing the train/test splits. However, this dataset is tailored for the *classification* scenario, that is each sample of the test split is guaranteed to belong to a fixed set of relation types. Datasets ReTACRED (Stoica et al., 2020) and Nyt (Riedel et al., 2010) include a special relation type for rejection evaluation, i.e., a specific output class that identifies couples of mentions that do not belong to the targeted relation types. However, these datasets do not include side-information about relation types, and can therefore not be straightforwardly used for zero-shot learning.

Unfortunately, to the best of our knowledge, no currently available RE dataset was initially built including both data for rejection evaluation and side-information required for zero-shot learning, except FewRel 2.0 (Gao et al., 2019). However, its authors' goal is to identify relations between mentions with respect to the whole set of types, whereas in our RE scenario, we aim to reject a candidate with respect to an "on-the-fly defined" list of targeted relations. Therefore, in this work, we choose to focus on FewRel and WikiZSL, with tailored evaluation procedures for rejection. This choice is motivated by the fact that we want to study *model robustness to adaptation for large scale and on-the-fly RE*: As such, we use the same dataset for evaluation with and without a rejection mechanism. These datasets remain relatively small and well-annotated compared to the data we will encounter in our scenario. We justify this choice by the need for comparability to previous models and by the ease of evaluation through reliable and exhaustive annotation at the sentence level.

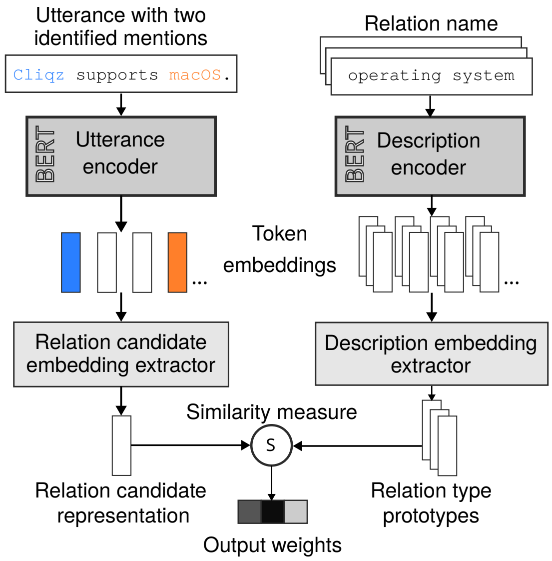

Figure 2: Generic illustration of an encoder-only zero-shot RE model.

2.3. Benchmarking Design Choices

Baldini Soares et al. (2019) and Thomas et al. (2023) compared different entity pair embedding methods, and conclude that surrounding the mentions of entity pairs with markup and concatenating the markup's embeddings offers the best performance for RE, unfortunately preventing offline encoding. This was also observed in ablation studies in Zhao et al. (2023); Li et al. (2024b).

In the biomedical domain, Sarrouti et al. (2022) and Naguib et al. (2024) compared encoder-only and encoder-decoder architectures for RE and named entity recognition, respectively. They both conclude that encoder-only architectures lead to better results while having fewer parameters.

3. Typology of Zero-Shot RE Models

We introduce a novel typology, summarized in Table 1, that allows to compare existing models for on-the-fly zero-shot RE with offline encoding, including the need for a rejection mechanism.[4]

[4]Recent works also studied generative language models (LM) for zero-shot RE (Chia et al., 2022; Li et al., 2023; Liu et al., 2024; Zhang et al., 2023, *inter alia*). These approaches (including more recent retrieval-augmented generation approaches) need to run a full forward pass on the entire LM for each input utterance, meaning that it is not possible to pre-compute and cache input representations for offline encoding. We do not consider aforementioned models in this work and focus solely on methods based on encoder-only architectures.

	ZS-BERT Chen and Li (2021)	LAVEENTAIL Sainz et al. (2021)	RCL Wang et al. (2022b)	MATCHPROMPT Wang et al. (2022a)	REMATCHING Zhao et al. (2023)	CL&CD Yang et al. (2024)	EMMA Li et al. (2024a)	ALIGNRE Li et al. (2024b)	CE-DA Zhang et al. (2025)	GLIREL Boylan et al. (2025)
Neural architecture										
SBERT	✓				✓			✓		
Single pass	✓									✓
Rej. mechanism		✓								
Late interaction	✓		✓	✓	✓	✓		✓	✓	
Side-information										
Name								✓	✓	✓
Description	✓			✓	✓	✓	✓	✓	✓	
Aliases								✓	✓	
Mention types					✓					
Experimental Results										
Public code	✓	✓	✓		✓		✓	✓		✓
F1	77.90		90.73		92.58	96.79	94.67	93.09	95.17	94.20

Table 1: Typology of zero-shot relation extraction models, ordered chronologically. Row SBERT indicates the use of SBERT to encode side-information of relation types. The F1 row contains results as reported by original authors on FEWREL using 5 unknown relation types, without rejection mechanism. (We do not include results for MATCHPROMPT and ALIGNRE as they have not been tested on this dataset) Note that there is also a late interaction variant of EMMA, but it achieves significantly lower F1 scores.

3.1. Encoders

A simplified illustration of a generic encoder-only neural architecture as used in few-shot and zero-shot learning is depicted in Figure 2: the neural network builds (1) a representation of the relation candidate between a head entity and a tail entity in an input utterance using the *utterance encoder* on the left side and (2) a representation of each targeted relation type, also called a prototype, from their respective side-information using the *description encoder* on the right side. To this end, relation mentions and descriptions tokens are contextually encoded into token embeddings, which are processed into a single vector of a relation mention or relation type (prototype). Then, prediction reduces to searching the best fit between the candidate representation and relation type prototypes, often by maximizing the cosine similarity.

We now describe the encoder properties highlighted by our typology. First, although all models rely on BERT (Devlin et al., 2019) or a similarly pre-trained encoder for the utterance encoder, they differ in the description encoder: some use the same network; others instead rely on the sentence embedding model of Reimers and Gurevych (2019), which we call SBERT in the following.

Second, the utterance and description encoders can be joined into a single encoder (e.g. including cross-attention between the input utterance and relation type side-information) or they may be independent. We say that independent encoders allow for *late interaction*, which is mandatory for on-the-fly zero-shot RE with offline encoding.

Natural Language Inference (NLI). There is a single exception to this generic architecture: LAVEENTAIL (Sainz et al., 2021), for label verbalization and entailment, which is based on a NLI model. The model takes as input the utterance (first sentence) and a relation type side-information (second sentence), and predicts if the two sentences are related by entailment, disregarding in this case contradiction or neutrality. High entailment probability means the input relation candidate is of this relation type.

3.2. Relation Type Representations

The following side-information about relation types has been used to build relation type representations:

1. *name*, e.g. "`operating system`;
2. *description*, e.g. "`operating system (OS) on which a software...`";
3. aliases, e.g. "`OS`", "`supported OS`";
4. expected head and tail mention type names, e.g. "`software`".

For the simplest case, Boylan et al. (2025) rely only on relation type names as a very short and dense source of information. Instead, Li et al. (2024a) choose to rely only on descriptions, whereas more involved approaches include name, description and aliases (Li et al., 2024b). Finally, Zhao et al. (2023) also incorporate head and tail mention type information.

3.3. Relation Candidate Representation

Most few- and zero-shot RE models encode the relation candidate in the input of the utterance encoder. For example, the first utterance and relation candidate in Figure 1 (left) can be encoded as

"`<E1>Cliqz</E1> supports the <E2>macOS</E2> operating system.`".

Such encoding breaks compatibility with our on-the-fly zero-shot setting. Indeed, in the offline encoding step (i.e., storing utterance representations of the whole targeted text database in advance), we do not know what mention types will be of interest to the user; therefore, we cannot pre-identify relation candidates at this stage (as head and tail mention types depend on targeted relation types).

Therefore, an important property we seek for the utterance encoder is that it is *single pass*: it takes as input the raw utterance "`Cliqz supports the macOS operating system.`", and the same token representations of this utterance can be used no matter the target head and tail mentions, which allows on-the-fly zero-shot RE with offline encoding.

3.4. Rejection Mechanism

LAVEENTAIL (Sainz et al., 2021) is the only model that includes a native rejection mechanism, based on an extra relation type for rejection whose description is "`E1 and E2 are not related`", where `E1` and `E2` are replaced head and tail mentions, respectively. If this rejection description obtains the highest entailment score among all descriptions for an utterance, said utterance is rejected.[5]

4. Efficient Relation Extraction without Rejection

Next, we explain how we adapt each considered model to our relation extraction scenario.

4.1. Selected Baselines

Table 1 shows that no model is tailored for our use case: there is only one model which is single pass, but it results in way lower F1 score than others, and none of the late interaction model includes a rejection mechanism.

For this study, we select two state-of-the art late interaction models whose code is publicly available: REMATCHING (Zhao et al., 2023) and ALIGNRE (Li et al., 2024b). Moreover, we also include the variant without late interaction of EMMA in the study, in order to understand how much performance drop when using late interaction. We could have included GLIREL instead, which also features early interaction, but we stuck with EMMA for its close resemblance to the two other models' architecture, which allows better comparison.

For the three models, we compare their *off-the-shelf* implementations, that is software as distributed by authors that we train ourselves, with our own adaptation that enables single pass inference.

4.2. Single Pass Adaptation

As described in Section 3.3, most models encode the relation candidate in the input utterance. In this section, we briefly describe our single pass variants of the chosen baselines. The main idea is that we give to the encoder the raw input utterance, and then use BERT's output contextual embeddings as representation of targeted head and tail entity mentions. With this approach, the whole unaltered input utterance is encoded, and at test-time any two mentions can be identified as relation head and tail candidates, by extracting their respective token embeddings. For the three baselines, we update the model so that the single pass variant is as similar as possible to the original implementation.

Let (s, e_1, e_2) be an input, where s is an input sentence, and e_1 (resp. e_2) is the span of the head (resp. tail) mention of the relation. We denote $d \in \mathbb{N}_{>0}$ the dimension of BERT outputs.

EMMA. We first build a representation $f_{\text{me.}}(s, e_1)$ (resp. $f_{\text{me.}}(s, e_2)$) for the head (resp. tail) mention, and $f_{\text{sent.}}(s)$ for the global utterance.[6] We test different strategies to build the representation of e_1 (similarly e_2), which are all based on the BERT model outputs, i.e., contextual embeddings:

- **first**: use the mention's first token embedding;
- **projection**: concatenate the first and last token embeddings of the mention, and then project this vector into dimension d;[7]
- **mean pooling**: average the first and last token embeddings of the mention;
- **max pooling**: compute element-wise maximum between first and last token embeddings.

Then, we build a relation candidate representation $f_{\text{rel.}}(s, e_1, e_2) \in \mathbb{R}^{3d}$ as follows:

$$f_{\text{rel.}}(s, e_1, e_2) = f_{\text{sent.}}(s) \oplus f_{\text{me.}}(s, e_1) \oplus f_{\text{me.}}(s, e_2),$$

where $\oplus$ denotes vector concatenation. The rest of the model follows the original implementation.

[5]There are older works such as (Levy et al., 2017) that include a rejection mechanism, but they are not included in our typology as their experimental results are far from current state-of-the-art.

[6]Which is simply the contextual embedding of the `[CLS]` token.

[7]Parameters of the projection are learned, and we use the same projection for e_1 and e_2.

	|T|	|D|	Dist. me.	Avg. len.
WikiZSL	113	94 383	77 623	24.85
FewRel	80	56 000	72 954	24.95

Table 2: Datasets statistics, where T is the set of relation types, D the set of relation quadruples including utterance, entities and relation type (s, e_1, e_2, t), and the last two columns give the number of distinct entity mentions and the average length of utterances, respectively.

AlignRE. For this model, the relation candidate representation $f_{\text{rel.}}(\boldsymbol{s}, e_1, e_2) \in \mathbb{R}^d$ is the mean of head and tail mention representations:

$$f_{\text{rel.}}(\boldsymbol{s}, e_1, e_2) = \frac{1}{2}(f_{\text{me.}}(\boldsymbol{s}, e_1) + f_{\text{me.}}(\boldsymbol{s}, e_2)),$$

and we test the same strategies as for Emma.

When trying our different strategies on AlignRE, we remove the custom prompt that the original model appends when using entity markers, i.e., "`The relation between [MASK] E1 and [MASK] E2 is [MASK]`", where E1 and E2 are replaced by the head and tail mentions, respectively.[8] Indeed, this extra input is not compatible with offline encoding as it pre-identifies targeted head and tail mentions. The rest of the model follows the original implementation.

ReMatching. The upgrade is similar to the one of Emma, but instead of concatenating the 3 representations to build $f_{\text{rel.}}(\boldsymbol{s}, e_1, e_2)$, they are kept separated to compute 3 cosine similarities that are then aggregated, as per original implementation.

4.3. Experiments

Prediction. Let T be the set of all relations annotated in a given dataset. Let $T' \subseteq T$ be a subset of relation types. Each model first computes a vector $\boldsymbol{w} \in \mathbb{R}^{T'}$ Then, the prediction $\widehat{t}$ is simply the relation type of maximum score: where w_t is output score associated with relation type $t \in T'$ for the input.[9]

$$\widehat{t} = \arg\max_{t \in T'} w_t.$$

Data. We evaluate on FewRel (Han et al., 2018) and WikiZSL (Chen and Li, 2021). Dataset statistics are given in Table 2.

[8]The original model would build relation representations by concatenating the three [MASK] tokens' representations from the prompt, whereas we use the previously described strategies.

[9]$\boldsymbol{w} \in \mathbb{R}^{T'}$ denotes the real-valued vectors indexed by elements of T'.

Hyperparameters. Models are trained for 5 epochs with a learning rate of 2×10^{-5} for Emma and 10^{-5} for AlignRE and RE-Matching, using AdamW (Loshchilov and Hutter, 2019). For AlignRE and RE-Matching, the SBert encoder is frozen. To obtain reliable results, all experiments are repeated 3 times with different random seeds, and we report average and standard deviation.

Evaluation metric. Note that in our evaluation setting, an input is a tuple $(\boldsymbol{s}, e_1, e_2)$, and an output is a relation type $t \in T_{\text{eval}}$, where T_{eval} is the evaluation set of relation types, unseen during training. As such, this evaluation setting reduces to a standard multi-classification setting, and we simply report the F1 score on the test set.

To better analyze robustness, we report F1 scores with different number of unknown classes in T_{eval}; in practice we set $|T_{\text{eval}}| \in \{5, 10, 15\}$. For a given set T_{eval}, the dataset is trivially split between train and evaluation depending on the gold annotated output for an input triple $(\boldsymbol{s}, e_1, e_2)$.

Analysis. Results are given in Table 3. Strategies like mean pooling, max pooling or first obtain lower scores by a few points in some cases, but models ReMatching or AlignRE are able to mitigate this difference in performance: while still worse than Emma with 5 unseen types, they are better with more unseen relation types and remain stable when substituting the entity markup by other embedding strategies. This shows that in this setting, late encoding does not necessarily hurt performances. Overall, the efficient ReMatching variants perform best compared to other efficient variants, but AlignRE remains competitive.

The concatenation of both first and last mention tokens embeddings results in the lowest scores. Emma is overall the most affected by the change of embedding strategy, going from the highest scores to scores comparable or worse to other models.

Overall, the evaluated single pass strategies yield acceptable scores with minimal degradation compared to the original approaches, and are therefore suitable to our use case.

5. Rejection Mechanism

In the previous section, we compared off-the-shelf versions of three zero-shot RE models with our updated efficient variants. Unfortunately, this evaluation setting is quite artificial as it assumes that each relation candidate is known to be of one of the targeted types T_{eval}. In practice, we often aim for relation *extraction*, where, for each relation candidate, we must decide if it belongs to a type in T_{eval} or not. In other words, we must have an option to *reject* a candidate.

	FewRel			WikiZSL		
	5	10	15	5	10	15
Emma						
Off-the-shelf	98.4 ±1.2	84.5 ±4.5	79.5 ±6.2	88.2 ±7.0	67.9 ±8.8	62.8 ±6.6
Our efficient variants						
↪ first	94.5 ±1.6	79.5 ±3.0	70.2 ±4.6	81.6 ±4.8	64.7 ±7.7	57.0 ±2.0
↪ projection	94.6 ±1.2	77.8 ±4.8	67.2 ±6.1	79.0 ±6.2	59.0 ±9.9	50.5 ±2.3
↪ max pooling	94.6 ±1.5	79.9 ±2.3	70.2 ±4.8	83.9 ±3.5	65.9 ±9.1	58.4 ±1.8
↪ mean pooling	94.6 ±1.4	80.0 ±2.5	70.9 ±5.3	82.6 ±3.4	65.5 ±8.9	57.8 ±3.7
Re-Matching						
Off-the-shelf	92.7 ±4.1	84.1 ±5.3	75.0 ±3.7	86.3 ±6.3	84.0 ±6.3	74.3 ±8.0
Our efficient variants						
↪ first	91.6 ±4.7	84.0 ±6.6	75.4 ±3.8	86.7 ±6.1	85.2 ±6.7	74.8 ±8.5
↪ projection	77.7 ±5.2	71.7 ±2.8	59.8 ±13.2	73.0 ±4.5	70.1 ±5.8	55.6 ±4.5
↪ max pooling	92.6 ±3.9	84.2 ±5.4	75.3 ±3.8	86.2 ±6.0	84.2 ±6.1	74.0 ±8.5
↪ mean pooling	91.9 ±4.4	83.7 ±6.1	75.3 ±4.8	86.8 ±6.0	85.1 ±6.6	75.1 ±8.7
AlignRE						
Off-the-shelf	90.7 ±0.2	84.9 ±1.8	73.4 ±5.1	78.2 ±4.3	73.8 ±6.0	65.2 ±7.7
Our efficient variants						
↪ first	89.5 ±4.4	86.9 ±0.8	72.7 ±4.0	80.1 ±4.1	74.1 ±2.9	68.1 ±4.2
↪ projection	51.4 ±10.9	55.2 ±5.4	39.7 ±4.7	76.7 ±6.5	44.3 ±3.2	46.5 ±4.6
↪ max pooling	90.0 ±3.5	87.6 ±1.3	73.5 ±4.7	77.9 ±6.5	72.4 ±3.7	67.4 ±4.7
↪ mean pooling	90.4 ±2.9	87.9 ±3.4	73.0 ±3.3	80.6 ±4.0	74.3 ±4.5	67.9 ±3.2

Table 3: Results in terms of macro F1, without rejection mechanism. We compare the off-the-shelf softwares with our efficient variants that allow on-the-fly zero-shot relation extraction with offline encoding.

To this end, we assume an extra set of relation types R such that $R \cap T = \emptyset$, where all relation types in R are considered as *reject types*. For a given input $(\boldsymbol{s}, e_1, e_2)$, the model now builds an augmented score vector $\overline{\boldsymbol{w}} \in \mathbb{R}^{T' \cup R}$, $T' \subseteq T$ s.t.:

$$\overline{w}_t = \begin{cases} w_t & \text{if } t \in T', \\ u_t & \text{otherwise,} \end{cases}$$

where $\boldsymbol{w}$ is built as described in Section 4, and $\boldsymbol{u} \in \mathbb{R}^R$ is a vector of weights for rejections. Then, prediction $\widehat{t}$ is:

$$\widehat{t} = \arg\max_{t \in T' \cup R} \overline{w}_t,$$

where no relation is predicted between e_1 and e_2 if $\widehat{t} \in R$. This can be performed by simply adding negative examples (irrelevant relation mentions) to the dataset, and labelling them with a negative type $t' \in R$. This setup, however, creates a strong class imbalance (the negative class is often prevalent) and does not guarantee a rejection mechanism fit for new unknown relation types, as we require. We describe 3 different rejection methods.

5.1. Proposed Methods

Rejection Threshold. The simplest mechanism learns a threshold weight such that a relation type can be predicted if and only if its weight is higher than this threshold (Sabo et al., 2021). This threshold is initially set to 0.5 and further learned as a parameter of the model.

In this setting, $R = \{r\}$ is singleton, and u_r is a learned parameter. It is easy to see that if $\forall t \in T' : w_t < u_r$, the input relation is rejected.

Rejection Description. Another possible mechanism consists in assuming the relation class is defined as any other class, that is we use a type description like "`There is no relation between the two entities.`" (Sainz et al., 2021; Thomas et al., 2024).

In this setting, $R = \{r\}$ and u_r is computed as other relation type scores.

Rejection Prototypes. The last approach we test is a variant of *multiple none of the above vectors* (Sabo et al., 2021). In this setting, $|R| > 1$, and we learn several reject type prototypes (or rejection relation type representations).[10] Then, u_r is the cosine similarity measure between the relation candidate representation and the reject type prototype for $r \in R$, computed in the same manner as other output weights. We set $|R| = 5$, as the

[10] This means that these vectors are learned parameters, and not outputs of a description encoder.

	FewRel			WikiZSL		
	5	10	15	5	10	15
Emma						
Threshold	9.8 ± 4.0	7.3 ± 8.9	9.8 ± 9.3	18.0 ± 6.4	28.7 ± 9.6	25.8 ± 12.7
Description	42.8 ± 16.6	42.5 ± 28.5	24.9 ± 15.3	61.4 ± 14.2	49.8 ± 16.0	47.3 ± 9.5
Prototypes	100.0 ± 0.0	99.8 ± 0.3	100.0 ± 0.0	37.1 ± 33.4	36.0 ± 54.5	67.5 ± 26.4
Re-Matching						
Threshold	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	84.3 ± 16.1	97.6 ± 4.1	90.8 ± 14.8
Description	99.3 ± 0.6	94.4 ± 7.1	86.9 ± 8.8	91.8 ± 11.5	82.1 ± 18.4	85.7 ± 10.8
Prototypes	95.6 ± 6.4	91.1 ± 6.3	81.5 ± 7.1	98.8 ± 0.9	81.0 ± 12.1	73.5 ± 4.4
AlignRE						
Threshold	100.0 ± 0.0	99.7 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	99.7 ± 0.0	99.8 ± 0.1
Description	86.3 ± 10.3	26.6 ± 10.5	27.5 ± 4.3	68.6 ± 14.0	34.3 ± 1.4	33.4 ± 3.9
Prototypes	99.1 ± 0.1	79.0 ± 10.0	65.5 ± 4.2	94.9 ± 2.3	63.9 ± 10.5	69.7 ± 3.6

Table 4: Rejection accuracy in the rejection pass.

	FewRel			WikiZSL		
	5	10	15	5	10	15
Emma						
Threshold	52.4 ± 1.9	31.8 ± 6.5	29.0 ± 11.0	54.6 ± 6.0	47.1 ± 8.4	47.0 ± 8.1
Description	41.3 ± 12.2	30.1 ± 4.2	20.6 ± 6.0	30.6 ± 8.5	24.8 ± 3.5	24.5 ± 6.0
Prototypes	0.0 ± 0.0	0.2 ± 0.3	0.0 ± 0.0	29.9 ± 13.8	25.8 ± 19.2	23.7 ± 22.9
Re-Matching						
Threshold	0.1 ± 0.2	0.7 ± 1.2	0.1 ± 0.2	26.7 ± 35.1	9.4 ± 3.1	26.7 ± 38.2
Description	30.8 ± 13.5	27.4 ± 19.7	24.7 ± 7.8	49.1 ± 34.0	31.2 ± 7.4	29.1 ± 14.0
Prototypes	44.3 ± 10.9	35.9 ± 14.4	39.5 ± 11.6	54.4 ± 1.1	28.9 ± 4.6	37.3 ± 12.5
AlignRE						
Threshold	0.0 ± 0.0	1.8 ± 3.0	0.0 ± 0.0	0.1 ± 0.1	9.4 ± 0.1	4.2 ± 3.3
Description	79.7 ± 3.4	66.0 ± 2.5	55.9 ± 1.6	64.5 ± 5.3	50.9 ± 8.3	51.1 ± 3.0
Prototypes	70.0 ± 8.0	71.9 ± 2.2	59.7 ± 1.3	46.6 ± 3.7	34.5 ± 5.6	43.4 ± 6.8

Table 5: Macro F1 measure in the retention pass, computed for (non reject) relation types.

original article said this value had little impact on performance when ranging from 1 to 20.

5.2. Training Loss

Training a zero-shot RE model with a rejection mechanism is challenging. First, every training instance is labeled with a training relation type. Second, in the case of rejection prototypes, we have several rejection candidates to train. Our loss is based on the squared hinge loss (Crammer and Singer, 2002), defined as follows:[11]

$$\ell_{\text{h2}}(\overline{\boldsymbol{w}}; t) = \max(0, 1 - \overline{w}_t + \max_{t' \neq t} \overline{w}_{t'})^2,$$

where $\overline{\boldsymbol{w}}$ is a vector of output scores and t is the index of gold class.

[11] In preliminary experiments, we tested training with the negative log-likelihood loss, but it consistently results in very low performance.

We propose a novel loss for training with a rejection mechanism that is built around a *ranking* objective: the gold relation type should be ranked above all other relation types (pos. vs. neg.), the gold relation type should be ranked above all rejection types (pos. vs. rej.), and one rejection type should be ranked above all non-gold types (rej. vs. neg.). Note that the last case is a partial labeling learning problem (Cour et al., 2011): the target may be a non singleton set R.

Intuitively, our approach learns to predict the gold relation type if it appears in the targeted types, otherwise we should predict any rejection type. This leads to the following aggregate loss:[12]

$$\ell_{\text{rej.}}(\overline{\boldsymbol{w}}; t) = \underbrace{\ell_{\text{h2}}\left([\overline{w}'_t]_{t' \in T'} ; t \right)}_{\text{pos. vs. neg.}}$$

[12] Following all baseline implementations, T' is the set of all gold classes that appear in the mini-batch.

$$+\underbrace{\ell_{\text{h2}}\left(\ [\overline{w}'_t]_{t' \in R \cup \{t\}}\ ; t\ \right)}_{\text{pos. vs. rej.}}$$

$$+\underbrace{\min_{r \in R} \ell_{\text{h2}}\left(\ [\overline{w}'_t]_{t' \in R \cup T' \setminus \{t\}}\ ; r\ \right)}_{\text{rej. vs. neg.}}.$$

The last term can be understood as an "hard EM" approach for partial labeling (Corro, 2024), where first search for the best rejection type, and then use it in a supervised loss. Such partial labeling losses are also referred to as "inf-losses" (Cabannes et al., 2020; Stewart et al., 2023).

5.3. Experiments

Evaluation protocol. To simulate rejection cases in datasets which feature no explicit annotation for that, we perform two passes on the evaluation split:

1. **Retention pass:** for each input, the model predicts one of the types in T_{eval} (i.e., model has to predict a relation type other than the rejection types in R);

2. **Rejection pass:** each input has its gold output type removed from the target set (model has to predict rejection).

Evaluation metrics. To measure the performance of each model, we evaluate them with respect to two metrics: the rejection accuracy in the rejection pass (Table 4), and the macro F1 measure on (non reject) relation types in the retention pass (Table 5).

Analysis. Note that a model that learns to reject every input will have a rejection accuracy of $100\,\%$. As such, when evaluating with a rejection mechanism, we search for a trade-off between the rejection accuracy and the F1 measure on relation types. For example, EMMA with the prototype rejection mechanism on FEWREL learns to reject everything.

The threshold strategy performs worse over all models: it pushes to learn either to reject almost everything or almost nothing.

ALIGNRE performs best overall using description and prototype rejection mechanisms: its rejection accuracy is high while maintaining good F1 in the retention pass. RE-MATCHING follows closely, especially on 10 or 15 unseen relation types.

6. Conclusion

In this work, we argued that most zero-shot RE works do not evaluate with real scenario in mind. We therefore introduce a challenging use case, and propose solutions to adapt several state-of-the-art models.

More specifically, we highlight two mandatory upgrades for efficient zero-shot RE: single pass adaptation and rejection mechanism augmentation. For both upgrades, we propose several strategies, and evaluate them on two datasets.

Acknowledgments

Experiments were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several universities (`https://www.grid5000.fr`). Hugo Thomas was funded by the "AI for semantic data analytics" doctoral program ANR-20-THIA-0018-01 and by Région Bretagne. Caio Corro is supported by the SEMIAMOR (CE23-2023-0005) and InExtenso (ANR-23-IAS1-0004) project grants of the French National Research Agency. This work contributes to the joint laboratory SYNAPSES (grant ANR-23-LCV2-0007) and to the AI Excellence Cluster SequoIA (grant ANR-23-IACL-0009).

Bibliographical References

Maristella Agosti, Giorgio Di Nunzio, Stefano Marchesin, Gianmaria Silvello, et al. 2018. A relation extraction approach for clinical decision support. In *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018); 12th International Workshop on Data and Text Mining in Biomedical Informatics (DTMBio 2018)*, volume 2482, Torino, Italy. CEUR Workshop Proceedings.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Marília Barandas, Duarte Folgado, Ricardo Santos, Raquel Simão, and Hugo Gamboa. 2022. Uncertainty-based rejection in machine learning: Implications for model development and interpretability. *Electronics*, 11(3):396.

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 (ACL 2022), pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Jack Boylan, Chris Hokamp, and Demian Gholipour Ghalandari. 2025. GLiREL - Generalist model for zero-shot relation extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*, pages 8230–8245, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Vivien Cabannes, Alessandro Rudi, and Francis Bach. 2020. Structured prediction with partial labelling through the infimum loss. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 1230–1239. PMLR.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*, pages 3470–3479, Online. Association for Computational Linguistics.

Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 345–356, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Proceedings of Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland. Association for Computational Linguistics.

Caio Corro. 2024. A fast and sound tagging method for discontinuous named-entity recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 19506–19518, Miami, Florida, USA. Association for Computational Linguistics.

Timothee Cour, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. *Journal of Machine Learning Research*, 12(42):1501–1536.

Koby Crammer and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292.

Shumin Deng, Yubo Ma, Ningyu Zhang, Yixin Cao, and Bryan Hooi. 2024. Information extraction in low-resource scenarios: Survey and perspective. In *Proceedings of the 15th IEEE International Conference on Knowledge Graphs (ICKG 2024)*, pages 33–49, Abu Dhabi, United Arab Emirates. IEEE.

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 (NAACL 2019)*, pages 4171–4186, Minneapolis, Minnesota, USA. 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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Hassan Hamad, Abhinav Kumar Thakur, Nijil Kolleri, Sujith Pulikodan, and Keith Chugg. 2024. FIRE: A dataset for financial relation extraction. In *Proceedings of Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3628–3642, Mexico City, Mexico. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Tingting Hang, Shuting Liu, Jun Feng, Hamza Djigal, and Jun Huang. 2025. Few-shot relation extraction based on prompt learning: A taxonomy, survey, challenges and future directions. *ACM Computing Surveys*, 58(2):40.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: A survey. *Machine Learning*, 113(5):3073–3110.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, pages 2410–2421, Toulon, France. ICLR.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, page 39–48, New York, NY, USA. Association for Computing Machinery.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. In *Proceeding of Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore. Association for Computational Linguistics.

Shilong Li, Ge Bai, Zhang Zhang, Ying Liu, Chenji Lu, Daichi Guo, Ruifang Liu, and Sun Yong. 2024a. Fusion makes perfection: An efficient multi-grained matching approach for zero-shot relation extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024)*, pages 79–85, Mexico City, Mexico. Association for Computational Linguistics.

Zehan Li, Fu Zhang, and Jingwei Cheng. 2024b. AlignRE: An encoding and semantic alignment approach for zero-shot relation extraction. In *Proceedings of Findings of the Association for Computational Linguistics: ACL 2024*, pages 2957–2966, Bangkok, Thailand. Association for Computational Linguistics.

Siyi Liu, Yang Li, Jiang Li, Shan Yang, and Yunshi Lan. 2024. Unleashing the power of large language models in zero-shot relation extraction via self-prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13147–13161, Miami, Florida, USA. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, Lousiana, USA. OpenReview.net.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.

Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. 2024. A survey on open information extraction from rule-based model to large language model. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9586–9608, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'10)*, pages 148–163, Barcelona, Spain. Springer-Verlag.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mourad Sarrouti, Carson Tao, and Yoann Mamy Randriamihaja. 2022. Comparing encoder-only and encoder-decoder transformers for relation extraction from biomedical texts: An empirical study on ten benchmark datasets. In *Proceedings of the 21st Workshop on Biomedical Language Processing (BioNLP 2021)*, pages 376–382, Dublin, Ireland. Association for Computational Linguistics.

Lawrence Stewart, Francis Bach, Felipe Llinares-Lopez, and Quentin Berthet. 2023. Differentiable clustering with perturbed spanning forests. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, volume 36, pages 31158–31176. Curran Associates, Inc.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2020. Re-TACRED: A new relation extraction dataset. In *Proceedings of the 4th Knowledge Representation and Reasoning Meets Machine Learning Workshop (KR2ML 2020, at NeurIPS'20)*, Online.

Hugo Thomas, Guillaume Gravier, and Pascale Sébillot. 2023. Derrière les plongements de relations. In *Proceedings of the 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2023)*, pages 311–322, Paris, France. ATALA.

Hugo Thomas, Guillaume Gravier, and Pascale Sébillot. 2024. One-shot relation retrieval in news archives: adapting n-way k-shot relation classification for efficient knowledge extraction. *Procedia Computer Science*, 246:1060–1069.

Jiaxin Wang, Lingling Zhang, Jun Liu, Xi Liang, Yujie Zhong, and Yaqiang Wu. 2022a. Match-Prompt: Prompt-based open relation extraction with semantic consistency guided clustering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 7875–7888, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. 2022b. RCL: Relation contrastive learning for zero-shot relation extraction. In *Proceedings of Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, Washington, USA. Association for Computational Linguistics.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691.

Zongqiang Yang, Junbo Fei, Zhen Tan, Jiuyang Tang, and Xiang Zhao. 2024. CL&CD: Contrastive learning and cluster description for zero-shot relation extraction. *Knowledge-Based Systems*, 293:111652.

Fu Zhang, He Liu, Zehan Li, and Jingwei Cheng. 2025. CE-DA: Custom embedding and dynamic aggregation for zero-shot relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 9814–9823, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada. Association for Computational Linguistics.

Jun Zhao, WenYu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng, and Mingming Sun. 2023. RE-Matching: A fine-grained semantic matching method for zero-shot relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 6680–6691, Toronto, Canada. Association for Computational Linguistics.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):293.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 5218–5230, Online. Association for Computational Linguistics.