Improving Zero-shot Relation Classification via Automatically-acquired Entailment Templates

Mahdi Rahimi and **Mihai Surdeanu**
Department of Computer Science
University of Arizona, Tucson, Arizona, USA
{marahimi, msurdeanu}@arizona.edu

Abstract

While fully supervised relation classification (RC) models perform well on large-scale datasets, their performance drops drastically in low-resource settings. As generating annotated examples are expensive, recent zero-shot methods have been proposed that reformulate RC into other NLP tasks for which supervision exists such as textual entailment. However, these methods rely on templates that are manually created which is costly and requires domain expertise. In this paper, we present a novel strategy for template generation for relation classification, which is based on adapting Harris' distributional similarity principle to templates encoded using contextualized representations. Further, we perform empirical evaluation of different strategies for combining the automatically acquired templates with manual templates. The experimental results on TACRED show that our approach not only performs better than the zero-shot RC methods that only use manual templates, but also that it achieves state-of-the-art performance for zero-shot TACRED at 64.3 F1 score.

1 Introduction

Relation classification (RC) identifies the relation that holds between two entities that co-occur in the same text. For example, given the sentence: *Jane's White House desk has a sign "Home on the Range"*, the relation between *Jane* and *White House* is `employee_of`. Beyond this simple example, RC is a critical NLP task with important applications to many domains such as intelligence (Doddington et al., 2004) and biomedical (Nédellec et al., 2013; Krallinger et al., 2017).

Recent directions mitigate the amount of supervision necessary for RC by taking advantage of the knowledge stored in large language models (LLMs). For example, Sainz et al. (2021) reformulated RC as an entailment task based on templates that are manually created as the verbalizations of relation labels. Then they are used to formulate a hypothesis that can be verified with an off-the-shelf LLM entailment engine. Other directions feed prompts that capture the definition of the task and examples into encoder-decoder language models (Han et al., 2022b). However, most of these directions tend to rely on templates/prompts that are *manually-created* by domain experts. This strategy has a potential high cost, and also runs the risk of inserting undesired biases in the data. Directions that focus on automatically learning prompts often produce prompts that are nonsensical to humans (Shin et al., 2020)

Our work aims to limit the above drawbacks of template-based approaches for RC. In particular, we expand the approach of Sainz et al. (2021) with explainable templates that are automatically acquired from a large textual collection. For template acquisition, we modify the BERT-Informed Rule Discovery (BIRD) algorithm (Rahimi and Surdeanu, 2022), which, given a seed template, automatically generates templates with similar meaning. BIRD encodes templates using contextualized representations generated by a transformer network (Devlin et al., 2019), and then uses a similarity measure to acquire similar templates based on an extension of the distributional similarity principle (Harris, 1954) to templates. For example, given a manual template `{subj} was founded by {obj}`, the template `{subj} was created by {obj}` is automatically created.

We use these automatically-acquired templates to expand the pool of manual templates used for RC, and show that this expansion yields statistically significant performance improvements.

The key contributions of this paper are:

- We introduce a novel strategy for template generation for RC, which is based on adapting Harris' distributional similarity principle to templates encoded using contextualized representations.

Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 187–195

- We perform an empirical analysis of different strategies for how to combine automatically-acquired templates with templates that were manually generated. The strategies include selection using entailment score, selection using BIRD's similarity score, and selection guided by the lowest entropy. All in all, our best combination obtains state-of-the-art performance for zero-shot TACRED, at 64.3 F1 score.

2 Related Work

Supervised Relation Classification. Most recent approaches for supervised relation classification use pre-trained language models such as models with self-supervised objectives, e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and GPT-3 (Brown et al., 2020), or adapt sequence-to-sequence models to the task, such as T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and GLM (Du et al., 2021). Prior to the advent of large language models supervised models needed a large amount of labeled data for training models from scratch (Kambhatla, 2004, Zeng et al., 2014). Recent approaches outperform traditional approaches by finetuning language models (Wu and He, 2019, Joshi et al., 2020, Yamada et al., 2020, Wang et al., 2021b, Lyu and Chen, 2021, Paolini et al., 2021, Wang et al., 2022a) or prompting (Han et al., 2022b, Han et al., 2022a, Zhang et al., 2023).

Our work is different in that we focus on the zero-shot scenario, a common situation in the real world, where a RC system must be developed rapidly without the luxury of training data.

Zero-shot Relation Classification. Getting accurate annotations for RC can be expensive because it is challenging for annotators to identify and come to a consensus on the structural information required. This poses an obstacle for RC models, which have traditionally depended on direct supervision from a sufficient end-task training data. Standard supervised models often perform poorly when dealing with low-resource situations (Schick and Schütze, 2021), highlighting the importance of developing methods that perform well in low-resource settings. As a result, several approaches have been proposed for relation classification with few training examples (Han et al., 2018, Gao et al., 2019, Baldini Soares et al., 2019, Sabo et al., 2021, Sainz et al., 2021). For the problem of zero-shot relation classification, Rocktäschel et al. (2015) and Demeester et al. (2016) proposed the use of logic rules. Wang et al. (2022b) used silver standard data cleaned by a class-aware clean data detection mechanism to train a textual entailment engine. In the literature, zero-shot RC has been reformulated as other tasks such as reading comprehension (Levy et al., 2017), textual entailment (Obamuyide and Vlachos, 2018, Sainz et al., 2021, Wang et al., 2022b), summarization (Lu et al., 2022), span-prediction (Cohen et al., 2020), question answering (Cetoli, 2020), triple generation (Wang et al., 2022a, Wang et al., 2021a), and prompting (Gong and Eldardiry, 2021).

Our work fits within this latter group. However, our contribution is that it mixes manual seed templates with explainable templates automatically acquired using the distributional similarity principle tailored for templates. Our results indicate that this simple strategy yields state-of-the-art performance on a popular zero-shot RC task.

3 Background

As mentioned, our zero-shot approach for relation classification relies on an entailment engine that is fed automatically-generated templates. In this section we discuss the building blocks necessary for this idea.

3.1 The Relation Classification Task

In the relation classification task we intend to classify a sentence with two marked entities into a predefined set of relations, or indicate that none of the relations hold between them (none-of-the-above or NOTA). Each input is a triple $x_i = (s, e_1, e_2)_i$ which consists of a sentence s with an ordered pair of two entities e_1 and e_2 (each entity is a span over s). The output $r \in R \cup \{\text{NOTA}\}$ indicates the two entities conform to one of the relations defined in the target set of $R = \{r_1, r_2, ..., r_N\}$ or none of the relations hold.

3.2 Zero-shot RC as an Entailment Task

Since our work is based on the works of Sainz et al. (2021) that reformulated RC as an entailment task on TACRED dataset, we provide an overview of their work. In a zero-shot setup no training examples are provided to the model. Therefore, the RC model must make predictions on relation instances without seeing any related data prior to that. By reformulating RC as an entailment task, it is possible to take advantage of the existing off-the-shelf entailment engines and use them as-is. Fig-

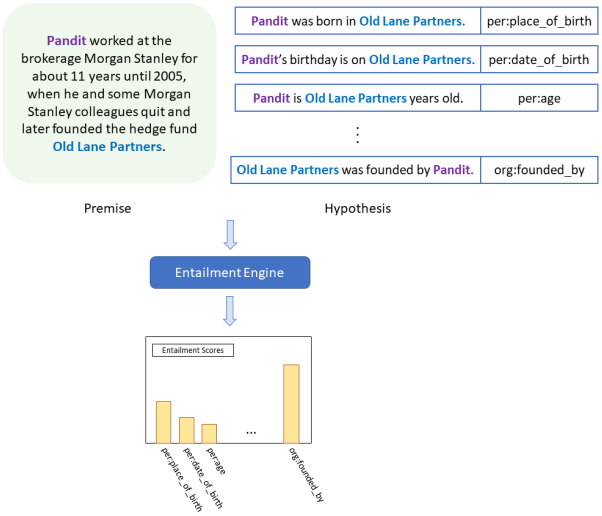

Figure 1: Overview of the entailment-based zero-shot relation classification approach via label verbalization.

ure 1 demonstrates the overview of the approach. First, a set of templates are manually created for each relation type in TACRED. The templates verbalize the relation types. For example, the relation `PER:DATE_OF_BIRTH` can be verbalized as `{subj}'s birthday is on {obj}`. Sainz et al. (2021) manually created 2 templates on average for each relation type. After that, an entailment engine is used to perform inference on each example in TACRED's test set. For each such example, the `{subj}` and `{obj}` placeholders of all the templates are replaced with the two marked entities of the example. After that, we use the TACRED example as premise and each template as hypothesis to feed them to an entailment engine. The entailment engine will produce an entailment score for each template. We pick the template with the highest score and return its corresponding relation type as the final prediction.

For the NOTA relation, however, creating templates does not produce good results. As a result, a threshold-based approach is used to detect NOTA. A threshold (between 0 and 1) is selected for the NOTA relation. If none of the entailment scores are above this threshold, the input example will be classified as NOTA. The treshold is selected by using 1% of TACRED's development set.

3.3 Rule Acquisition

BERT-Informed Rule Discovery or BIRD (Rahimi and Surdeanu, 2022) is a rule[1] acquisition algorithm. Informally, BIRD learns inferences from text such as *"X is the author of Y ≈ X writes Y"*. Some of these inferences are not exact paraphrases (but are still relevant and potentially useful) such as *"X is the author of Y ≈ X is known for Y"*. More formally, BIRD is initialized with a seed rule (e.g. *"X is the author of Y"*), which is implemented as a syntactic path connecting two concepts, and infers one or more possible matches (e.g. *"X writes Y"*) where each match is a syntactic path. BIRD generates the matches by implementing Harris' Distributional Hypothesis principle (Harris, 1954) to rules. It states that if two patterns tend to link the same sets of words, they tend to have similar meanings.

BIRD relies on contextualized representations generated by a transformer network (Devlin et al., 2019). In particular, a pattern has two slots (*X* and *Y*); BIRD computes contextualized embeddings for each slot. By doing so, a pattern will be represented by two embedding vectors. Then a (cosine) similarity score is calculated between each slot of any given two patterns, i.e., one similarity score is calculated for slot X and one similarity score is calculated for slot Y. Finally, the average of the two similarity scores is computed as the similarity score between the two patterns.

Given a corpus, BIRD extracts and stores all patterns from the corpus and then computes the contextualized embeddings for each slot of each pattern. Therefore, given a pattern, BIRD can search the extracted patterns space and find the most similar patterns to the input pattern.

3.4 TACRED Dataset

TACRED (Zhang et al., 2017) is a large-scale relation classification benchmark that is consisted of 106,264 examples and 42 relation types including the `no_relation (NOTA)` label. Each example contains the information about the entity type, among other linguistic information. TACRED examples include 68,124 for training, 22,631 for validation, and 15,509 for testing.

4 Approach

One limitation of (Sainz et al., 2021) is that they rely on *manually-generated* templates, which require effort to create and may prone to bias. Our work mitigates this limitation by automatically expanding the manual patterns using BIRD.

We utilize the manual templates as seed patterns to feed them into BIRD to generate new patterns that tend to have similar meanings to the

[1]Alternatively called patterns or syntactic paths.

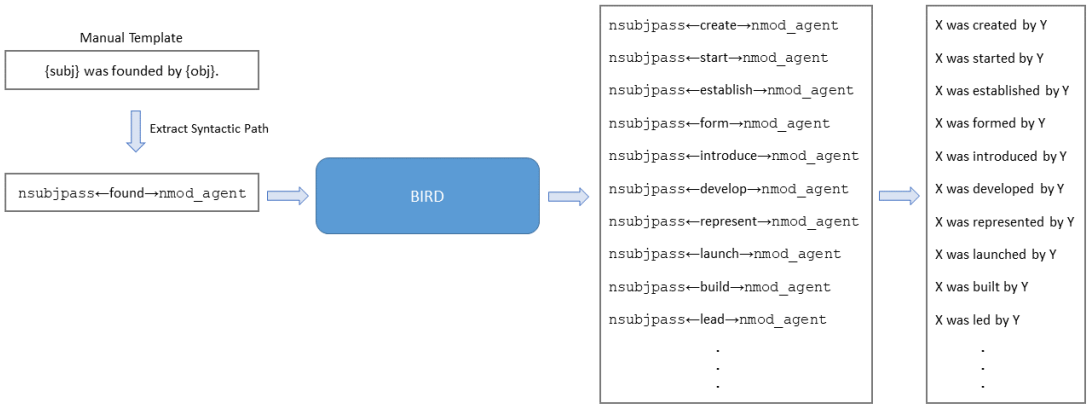

Figure 2: The process of using manual templates as seed pattern to generate similar patterns by BIRD.

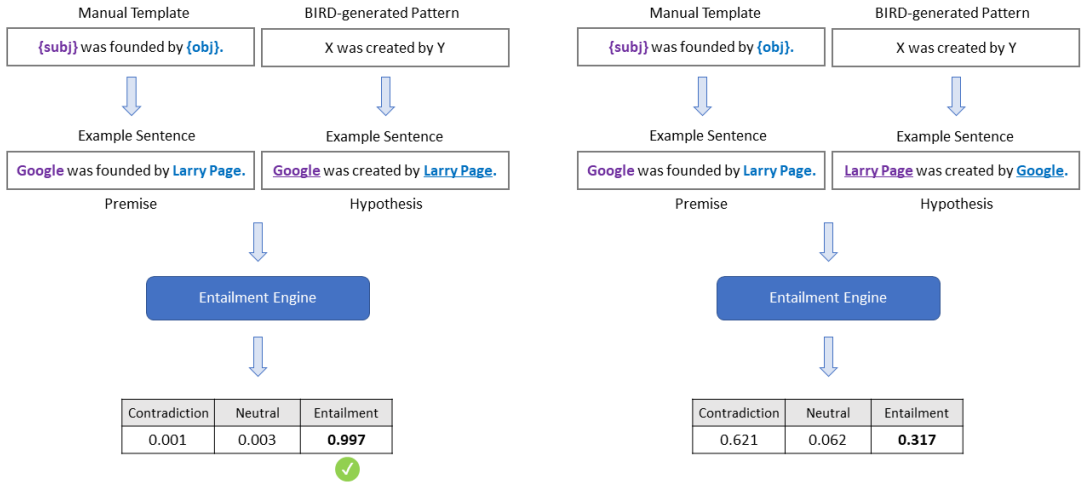

Figure 3: The process of determining the subject and object of a pattern generated by BIRD.

manual templates. Figure 2 shows this process. For each manual template, we extract its syntactic path. For example, for the template `{subj} was founded by {obj}`, we extract the syntactic path `nsubjpass←found→nmod_agent` (X was founded by Y). Then the syntactic path is fed to BIRD to generate new syntactic paths. We generate 40 patterns for each manual template. After the syntactic paths are generated, we generate their corresponding English text using a module we created. For example, for the syntactic path `nsubjpass←create→nmod_agent`, the English text "`X was created by Y`" is generated.

The corresponding patterns of 12 manual templates did not exist in BIRD's pattern collection. Therefore, it was not possible to use them as seed patterns. For these templates we manually selected an existing high-quality pattern that was very close in meaning. For example, the pattern *"X is spouse of Y"* did not exist in BIRD's pattern collection. We chose the pattern *"X married Y"* instead.

4.1 Determining Subject and Object of a Pattern

The manual templates have `{subj}` and `{obj}` placeholders which correspond to the subject and object of a relation, respectively. The patterns generated by BIRD do not contain this information. That is, it is unclear what the mapping is between the X and Y slots in a BIRD template and `{subj}` and `{obj}` in a relation template. Therefore, it is necessary to determine which slot of a pattern is the subject and which slot is the object. We use our off-the-shelf entailment engine to determine this information. Figure 3 demonstrates this procedure. For each generated pattern, we create an example sentence for its seed manual template. This example sentence will be used as premise. As for the hypothesis, we create two example sentences for the pattern by using the same words that filled the

placeholders of the seed manual template to fill the slots of the pattern. The two example sentences are the same sentence except their slot-filler words are swapped. One of these two hypotheses is expected to have a similar meaning to the manual template's example sentence. Therefore, we expect the entailment engine to produce a high entailment score for it and to produce a low score for the other one. We pick the hypothesis with the highest entailment score to determine the subject and object of the pattern. If the entailment engine produces entailment scores below 0.5 for both of the hypotheses, we discard the pattern. By determining the subject and object of the patterns, they become fully functioning templates.

4.2 Selection of Patterns

After the templates are generated and set up, a selection mechanism is required to keep the most useful ones and discard the rest.[2] We always keep all of the manual templates and the selection mechanism is only applied on the BIRD-generated ones.

4.2.1 Selection using entailment score

In subsection 4.1, it was explained that an entailment score is obtained for each BIRD-generated pattern when the subject and object of the pattern is determined. We use this entailment score to sort the templates and keep the top ones.

4.2.2 Selection using BIRD's similarity score

For a seed pattern, BIRD generates similar patterns using a similarity measure. This yields a similarity score for each BIRD-generated pattern. We use this similarity score to sort the templates and keep the top ones.

4.2.3 Selection guided by lowest entropy

A low entropy for the prediction of a classifier means that the classifier was more confident when making the prediction. We use this idea to pick the templates that together result in the lowest entropy. For each TACRED test set example, we perform the following. For each relation $r_i \in R$, we pick the template (from the pool of templates for r_i) that produces the highest prediction score (entailment score when TACRED example is premise and the template is hypothesis). For all $r_j \in R - \{r_i\}$, we pick the template (from the pool of templates for r_j) that produces the lowest prediction score. The combination of these chosen templates create one candidate set of templates. This procedure is done for each $r_i \in R$. As a result, we will have 41 candidate sets. Each candidate set is expected to have a low entropy. We keep the candidate set with the lowest entropy. Note that during the candidate sets creation process, we never look at the gold labels and merely use the model prediction scores.

4.3 Inference

After the final set of templates is selected, we perform the inference according to the method described in Sainz et al. (2021) as shown in Figure 1. During inference, they pick the template that produces the highest entailment score and return its corresponding relation type as the final prediction. In addition to using the "highest entailment score", we also experimented with two additional methods to perform the final prediction. Firstly, instead of choosing the relation type that has the highest entailment score, we computed the average of the top 3 highest entailment scores for each relation type, and then we pick the relation type that has the highest amount of this average. Secondly, we used a normalized unweighted voting mechanism. For each template, if its entailment score is more than its neutral and contradiction scores, we count this template as a positive vote. We count these positive votes for each relation type and then divide the count by the total number of templates for the relation type. The relation type with the highest normalized votes will be selected as the final prediction.

5 Experiments

5.1 Experimental Settings

We conducted our experiments on TACRED using the zero-shot RC setup from Sainz et al. (2021). For our experiments, we used DeBERTa which was the off-the-shelf entailment engine that produced the highest results in Sainz et al. (2021).

The set of templates that we used for our experiments included all of the manual templates as well as a small subset of the BIRD-generated templates. The subset was selected according to the pattern selection methods explained in subsection 4.2. Initially, we generate 40 patterns per each manual template. After that, we sort the generated patterns according to a scoring measure and keep the top k patterns with $k = \{1, 2, 3, 4\}$[3].

[2] Keeping all of the templates resulted in poor performance in our experiments.

[3] Values of k higher than 4 did not yield better performance.

The no_relation threshold was selected using 1% of the validation set. Sainz et al. (2021) divided the validation set to 100 stratified folds, ran 100 experiments for each fold, and obtained 100 f1 scores. They reported the median f1 score along with f1 standard deviation. We do the same.

5.2 Results

Table 1 shows the zero-shot RC results including our best results which were obtained using "entailment score" as pattern selection method and "highest entailment score" as inference method. Third-party results are as reported by authors. The table indicates that our results match the current state-of-the-art for zero-shot TACRED (Zhang et al., 2017), and are 1.5 F1 points above Sainz et al. (2021)'s results. The improvement in F1 comes from both better precision and better recall. In order to confirm that the difference in F1 scores between our method and (Sainz et al., 2021) is statistically significant, we ran a significance test using bootstrap resampling experiment with 1000 samples. This test indicated that the difference in F1 scores between our method and (Sainz et al., 2021) is statistically significant with a p-value of 0.3. We find these results encouraging considering that the manual patterns in Sainz et al. (2021) were manually developed by domain experts, whereas the pattern acquisition method in BIRD was developed independently of the TACRED task and its distributional similarity statistics were acquired from a different dataset (a Wikipedia subset) (Rahimi and Surdeanu, 2022).

We experimented with different pattern selection methods as shown in Figure 4. We observed that entailment score performed better overall than the rest. We also observed that as the number of added templates increased, the performance of the models decreased indicating that increasing the number of templates had a negative effect on performance. In general, while BIRD's similarity score was more robust as more templates are considered, the best result was obtained with one added template. Furthermore, we experimented with different inference methods as shown in Figure 5 when the pattern selection method was entailment score (since it produced the highest results). We observed that highest entailment score performed better than the rest. We hypothesize that this is caused by the relatively noisy BIRD generated templates, which suggest that future work is needed to better align BIRD's output with the entailment-driven RC task.

Model	Pr.	Rec.	F1
SuRE	-	-	20.6
DEEPSTRUCT	-	-	36.1
DEEPEX	-	-	49.2
Zero-shot SQuAD	49.7	78.9	57.1
$\text{NLI}_{\text{DeBERTa}}$	66.3	59.7	62.8 ± 1.7
Wang et al. (2022b)	-	-	**64.3 ± 1.2**
NLI_{BIRD} (ours)	68.8	60.4	**64.3 ± 1.3**

Table 1: Zero-shot RC results. Our best results was obtained by using "entailment score" as pattern selection method and "highest entailment score" as inference method. Top six rows are from third-party zero-shot RC systems as reported by authors. Third party results are from SuRE (Lu et al., 2022), DEEPSTRUCT (Wang et al., 2022a), DEEPEX (Wang et al., 2021a), Zero-shot SQuAD (Cohen et al., 2020), $\text{NLI}_{\text{DeBERTa}}$ (Sainz et al., 2021), and Wang et al. (2022b).

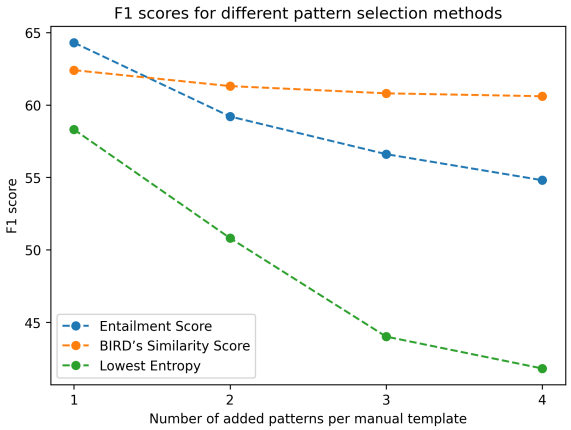

Figure 4: F1 scores for different pattern selection methods.

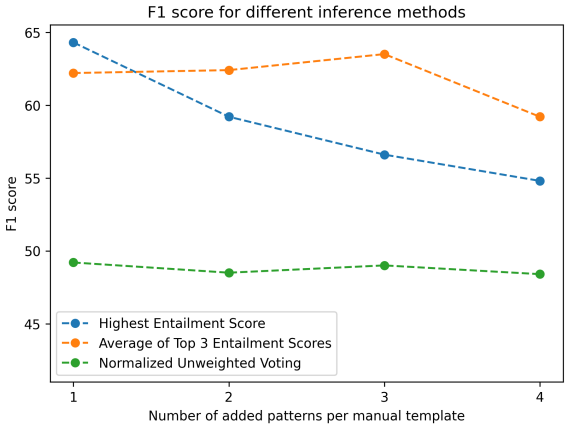

Figure 5: F1 scores for different inference methods when entailment score is used for pattern selection.

6 Conclusions

This paper introduces a zero-shot approach for relation classification. Our direction expands the method of Sainz et al. (2021), which converts the relation extraction task into textual entailment that is informed by manual templates that characterize the relations of interest. Unlike Sainz et al. (2021), we combine manual templates with templates that were automatically-acquired using an adaption of Harris' distributional similarity principle to templates encoded using contextualized representations.

We empirically evaluate our approach on a zero-shot setting of the TACRED relation classification task (Zhang et al., 2017). We investigated multiple strategies to rank the quality of the automatically-acquired templates. All in all, we found that a simple strategy, which considers the top template as ranked by a textual entailment engine, performs the best. Our results match the current state-of-the-art for zero-shot TACRED. We find this result exciting, especially considering that the template acquisition component is disconnected from the TACRED dataset. Beyond these results, this paper opens interesting questions for future work such as how to increase the relevance of the automatically-acquired templates for a specific task.

References

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*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alberto Cetoli. 2020. Exploring the zero-shot limit of FewRel. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1447–1451, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation classification as two-way span-prediction. *arXiv preprint arXiv:2010.04829*.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted rule injection for relation embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1389–1399, Austin, Texas. Association for Computational Linguistics.

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

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All nlp tasks are generation tasks: A general pretraining framework. *arXiv preprint arXiv:2103.10360*.

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)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Jiaying Gong and Hoda Eldardiry. 2021. Prompt-based zero-shot relation classification with semantic knowledge augmentation. *arXiv preprint arXiv:2112.04539*.

Jiale Han, Shuai Zhao, Bo Cheng, Shengkun Ma, and Wei Lu. 2022a. Generative prompt tuning for relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3170–3185, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022b. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

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*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181, Barcelona, Spain. Association for Computational Linguistics.

Martin Krallinger, Martin Pérez-Pérez, Gael Pérez-Rodríguez, Aitor Blanco-Míguez, Florentino Fdez-Riverola, Salvador Capella-Gutierrez, Anália Lourenço, and Alfonso Valencia. 2017. The biocreative v. 5 evaluation workshop: tasks, organization, sessions and topics.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 390–395, Online. Association for Computational Linguistics.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria. Association for Computational Linguistics.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*.

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

Mahdi Rahimi and Mihai Surdeanu. 2022. Do transformer networks improve the discovery of rules from text? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3706–3714, Marseille, France. European Language Resources Association.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1119–1129, Denver, Colorado. 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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2021a. Zero-shot information extraction as a unified text-to-triple translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1225–1238, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022a. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021b. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Tianyin Wang, Jianwei Wang, and Ziqian Zeng. 2022b. Learning with silver standard data for zero-shot relation extraction. *arXiv preprint arXiv:2211.13883*.

Shanchan Wu and Yifan He. 2019. Enriching pretrained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Wenjie Zhang, Xiaoning Song, Zhenhua Feng, Tianyang Xu, and Xiaojun Wu. 2023. Labelprompt: Effective prompt-based learning for relation classification. *arXiv preprint arXiv:2302.08068*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.