applied sciences

MDPI

Article

Commonsense Knowledge-Aware Prompt Tuning for Few-Shot NOTA Relation Classification

Bo Lv [1,2,3], Li Jin [1,2,*], Yanan Zhang [1,2,3], Hao Wang [4], Xiaoyu Li [1,2] and Zhi Guo [1,2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; lvbo19@mails.ucas.ac.cn (B.L.); zhangyanan161@mails.ucas.ac.cn (Y.Z.); lixy01@aircas.ac.cn (X.L.); guozhi@aircas.ac.cn (Z.G.)
[2] Key Laboratory of Network Information System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Information Science and Technology, North China University of Technology, Beijing 100190, China; wanghaomails@gmail.com
* Correspondence: jinlimails@gmail.com

Abstract: Compared with the traditional few-shot task, the few-shot none-of-the-above (NOTA) relation classification focuses on the realistic scenario of few-shot learning, in which a test instance might not belong to any of the target categories. This undoubtedly increases the task's difficulty because given only a few support samples, this cannot represent the distribution of NOTA categories in space. The model needs to make full use of the syntactic information and word meaning information learned in the pre-training stage to distinguish the NOTA category and the support sample category in the embedding space. However, previous fine-tuning methods mainly focus on optimizing the extra classifiers (on top of pre-trained language models (PLMs)) and neglect the connection between pre-training objectives and downstream tasks. In this paper, we propose the commonsense knowledge-aware prompt tuning (CKPT) method for a few-shot NOTA relation classification task. First, a simple and effective prompt-learning method is developed by constructing relation-oriented templates, which can further stimulate the rich knowledge distributed in PLMs to better serve downstream tasks. Second, external knowledge is incorporated into the model by a label-extension operation, which forms knowledgeable prompt tuning to improve and stabilize prompt tuning. Third, to distinguish the NOTA pairs and positive pairs in embedding space more accurately, a learned scoring strategy is proposed, which introduces a learned threshold classification function and improves the loss function by adding a new term focused on NOTA identification. Experiments on two widely used benchmarks (FewRel 2.0 and Few-shot TACRED) show that our method is a simple and effective framework, and a new state of the art is established in the few-shot classification field.

Keywords: commonsense knowledge-aware prompt tuning; few-shot none-of-the-above relation classification; pre-trained language models; scoring strategy

Citation: Lv, B.; Jin, L.; Zhang, Y.; Wang, H.; Li, X.; Guo, Z. Commonsense Knowledge-Aware Prompt Tuning for Few-Shot NOTA Relation Classification. *Appl. Sci.* **2022**, *12*, 2185. https://doi.org/10.3390/app12042185

Academic Editors: Andrzej Sobecki, Higinio Mora, Doina Logofătu and Julian Szymanski

Received: 24 November 2021
Accepted: 14 February 2022
Published: 19 February 2022

1. Introduction

In recent years, few-shot none-of-the-above relation classification has received widespread attention due to the fact that it is more in line with real-world applications. In the original *N*-way *K*-shot relation classification, all queries are assumed to be in the given relations set. However, the vast majority of sentences do not express specific relations or relations that are in the given set, which should also be taken into consideration. This calls for the none-of-the-above (NOTA) relation, which indicates that the query instance does not express any of the given relations. As shown in Figure 1, the relation between two entities contained in the query instance does not belong to category A, B, or C. The model needs to recognize that there is no relationship between the two entities, so we choose D. It is very

difficult to classify the query by calculating the similarity of the query and support samples, especially for selecting the threshold that distinguishes the NOTA class from others.

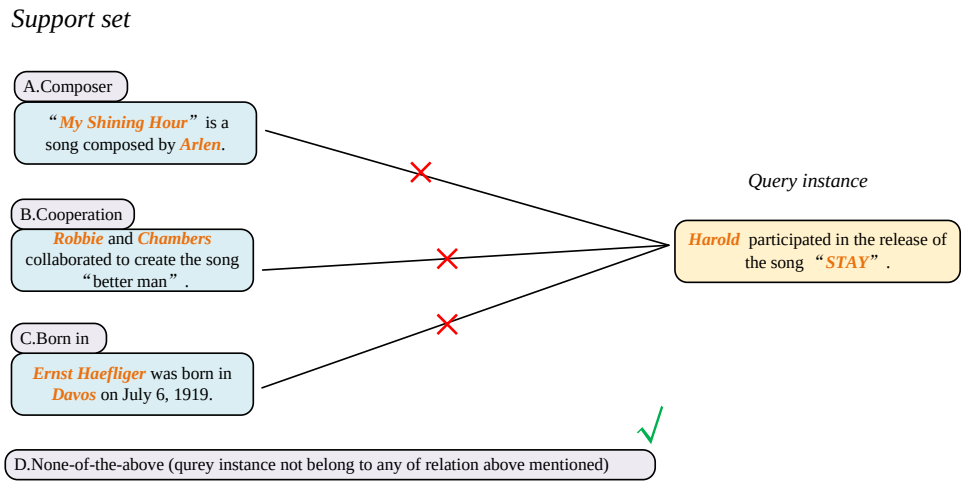


Figure 1. An example for a 3-way 1-shot scenario on few-shot NOTA relation classification. Calculate the similarity between query and each support sample. If the highest similarity value is greater than the NOTA category threshold, the query relation is the same with the support instance, which is most similar to the query instance; otherwise, the query relation is NOTA.

A lot of works have been devoted to identifying the NOTA relation. Han et al. [1] proposed a model named BERT-PAIR based on the sequence classification model in BERT [2], which treated NOTA the same as other relations and optimized the model with the cross-entropy loss. Ofer et al. [3] proposed a novel classification scheme in which the NOTA category threshold was represented as learned vectors in the embedding space. Nevertheless, there are still several non-trivial challenges for the few-shot NOTA relation classification. On the one hand, previous fine-tuning methods require adding extra classifiers on top of pre-trained language models (PLMs) and further training the models under classification objectives, which do not take full advantage of the knowledge learned during the pre-trained phase, especially when there is a NOTA relation between entities. On the other hand, the distances of negative pairs in the embedding space are loose. The score values of negative pairs after the softmax function are very small, which causes the model to learn about negative pairs insufficiently.

To address the limitation of current few-shot methods, we propose a commonsense knowledge-aware prompt tuning (CKPT) method for few-shot NOTA relation classification. First, we follow the route of prompt-based prediction developed by the GPT series [4], and introduce it into a few-shot NOTA relation classification. Prompt-based prediction treats the downstream task as a (masked) language modeling problem, where the model directly generates a textual response (referred to as a label word) to a given prompt defined by a task specific template (see Figure 2). Compared with conventional fine-tuning methods, prompt learning does not require extra neural layers and closes the objective formal gap between pre-training and fine-tuning. Second, we use external commonsense knowledge to generate a set of expanded label words for each original label, which are not only synonyms, but also cover different granularities and perspectives. These expanded labels are more comprehensive and unbiased expressions for original relation labels. For example, the naive verbalizer *none* means that only predicting the word *none* is regarded as a NOTA relation during inference, regardless of the predictions of other relevant words, such as *without* and *nor*, which are also informative for NOTA relation. Third, we propose a NOTA loss function to optimize the similarity score on recognizing the NOTA relation, which improves the problem that the score values of negative pairs after the softmax function are too small. When training our backbone, the NOTA loss function is added to the overall

loss to encourage the model to accurately detect NOTA examples, in addition to accurately performing the episode's classification task.

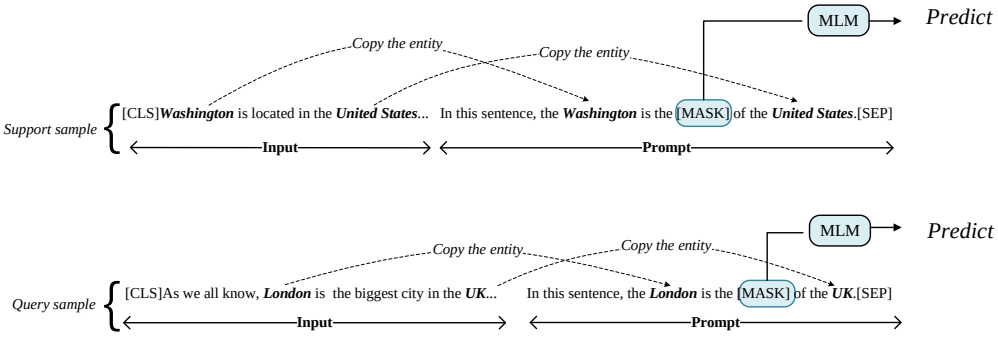


Figure 2. The illustration of prompt tuning for obtaining the vector (mask) prediction of the support sample and query sample. MLM is the self-supervised masked language model of the BERT.

The main contributions of our paper can be summarized as follows: (1) We propose a commonsense knowledge-aware prompt tuning model for the few-shot NOTA relation classification that injects commonsense knowledge into prompt label construction. (2) We design a learned scoring strategy on top of the embedding of our model, which can distinguish the NOTA pairs and positive pairs in embedding space more accurately. (3) Extensive experiments on two few-shot benchmark datasets (FewRel 2.0 and Few-shot TACRED) illustrate the effectiveness of our model in low resource NOTA relation settings.

2. Related Work

2.1. Few-Shot Relation Classification

The few-shot relation classification [5] task aims to classify the semantic relation under a few annotated data [6] of domain relation classes. Han et al. [7] first proposed the FewRel dataset for few-shot relation classification, and adopted some state-of-art few-shot methods intended for computer vision, including meta networks [8], few-shot GNN [9], and neural attentive meta-learning [10] to the FewRel dataset.

Meta-learning [11] is the science of systematically observing how different machine learning approaches perform on a wide range of learning tasks, and then learning from this experience, or meta-data, to learn new tasks much faster than would otherwise be possible. Specifically, it samples few-shot classification tasks from training samples belonging to the base classes and then optimizes the model to perform well. The meta-learning based methods can be roughly categorized into two groups (memory-based methods and optimization-based methods). Memory-based methods are based on the idea of training a meta-learner with memory to learn novel concepts. Meta-LSTM [12] trained an LSTM-based meta-learner to learn the exact optimization algorithm, as well as a mechanism for updating the learner's parameters by a handful of the sample set. Similarly, Meta-SGD [13] trained a meta-learner that can produce learner's initialization in just one step, on both supervised learning and reinforcement learning. To improve the performance of the model with less training data, MICK [14] aggregated cross-domain knowledge into models by open-source task enrichment. The model aimed to classify query instances, and sought basic knowledge about supporting examples to obtain a better example representation. Wang et al. [15] proposed the CTEG model, which is trained by entity-guided attention and is confusion-aware to decouple easily confused relations. Optimization-based methods follow the idea of differentiating an optimization process over support-set within the meta-learning framework. Dong et al. [16] proposed a novel meta-information-guided few-shot relation classification model (MAML) which utilizes semantic concepts of classes, guiding meta-learning in both initialization and adaptation. To handle the uncertainty of the prototype vectors, Qu et al. [17] used the stochastic gradient Langevin dynamics

(SGLD), which parameterized the initial prior of the prototype vectors with a graph neural network on the global relation graph.

Another branch in the few-shot [18] relation classification field is metric-learning-based methods, which embed the samples into a metric space so that the samples can be classified according to similarity to or distance between each other. Relation network [19] adapted the convolutional neural network to extract the features of support and query samples, and the relation classification scores were obtained by concatenating the vectors of support and query samples into the relation network. To overcome the catastrophic forgetting problem, Cai et al. [20] introduced a two-phase prototypical network, which adapted prototype attention alignment and triplet loss to dynamically recognize the novel relations with a few support instances without catastrophic forgetting. Similarly, Fan et al. [21] proposed the large-margin prototypical network with fine-grained features (LM-ProtoNet), which could generalize well on few-shot relations classification. To learn predictive and robust relation representations from the training phase, Ding et al. [22] proposed prototype learning methods with geometric interpretation, where the prototypes were unit vectors uniformly dispersed in a unit ball, and the sentence embeddings were centered at the end of their corresponding prototype vectors. Wu et al. [23] expanded the mean selection to dynamic prototype selection by fusing a self-attention mechanism and proposed a query-attention mechanism to more accurately select prototypes. Our approach is based on a pre-trained encoder [2], which belongs to a metric-learning method.

2.2. Open-World Detection

The essence of the NOTA category resembles open-world detection, as in both cases, the goal is to detect instances not falling under the known categories. Tan et al. [24] defined the OOD classes as the set of all classes that were not part of the training classes (vs. NOTA, which means that none of the given support classes in an episode are present). Andreas L. et al. [25] proposed a novel framework as a solution to the open world learning problem. Willes et al. [26] proposed the small-context and large-context few-shot open-world recognition (FS-OWR) problem settings, extending the scope of the existing open-world recognition setting to include learning with limited labeled data. The above work made great progress in image open-domain recognition, but there are few studies on open-domain few-shot relation classification tasks.

2.3. Prompt-Tuning

Since the emergence of GPT3 [27], prompt tuning has received considerable attention. GPT-3 [27] demonstrates that with prompt tuning and in context learning, large-scale language models can achieve superior performance in the low-data regime. The authors of Ref. [27] suggest that this framework is powerful and attractive for a number of reasons: it allows the language model to be pre-trained on massive amounts of raw text, and by defining a new prompting function, the model is able to perform few-shot or even zero-shot learning, adapting to new scenarios with few or no labeled data. Liu et al. [28] surveyed and organized research works in a new paradigm in natural language processing, which was dubbed "prompt-based learning". They introduced the basics of this promising paradigm, described a unified set of mathematical notations that could cover a wide variety of existing work, and organized existing work along several dimensions, e.g., the choice of pre-trained models, prompts, and tuning strategies.

The following works [29,30] argued that small-scale language models [2,31,32] could also achieve decent performance using prompt tuning. Some research works have been conducted on text classification or the tasks in SuperGLUE [33]. Ding et al. [34] applied prompt tuning to entity typing with prompt learning by constructing an entity-oriented verbalizer and templates. To avoid label-intensive prompt design, automatic searches for discrete prompts have been extensively explored. Gao, Fisch, and Chen et al. [35] first explored the automatic generation of label words and templates. Shin et al. [36] designed automatic verbalizer searching methods for better verbalizer choices. However, their

methods required an adequate training set and validation set for optimization. Recently, some continuous prompts have also been proposed [37,38], which directly utilize learnable continuous embeddings as prompt templates. For relation extraction, Han et al. [30] proposed a model called PTR, which applied logic rules to construct prompts with several sub-prompts. Previous fine-tuning methods mainly focus on optimizing additional classifiers, thus requiring more training samples to converge. Prompt methods reformulate downstream tasks as closed tasks with textual templates and a set of label words, and the design of templates is proved to be significant for prompt-based learning. In this work, we propose the CKPT model, which uses external knowledge to boost the performance of prompt tuning. Compared to the previous strategies, our method can effectively utilize more than 50 related label words from common knowledge for each class, and can be effectively applied to a few-shot NOTA relation classification.

3. Materials and Methods

This section introduces the overall framework of our CKPT model for NOTA few-shot relation classification. The overall architecture of our CKPT model is shown in Figure 3. In the following, we first introduce the task definition (Section 3.1), and then give the details of our proposed method: (1) the commonsense knowledge-aware prompt tuning method that measures the distance in the learned embedding of a few-shot classifier (Sections 3.2 and 3.3), and (2) the learned scoring strategy on top of the embedding of CKPT, which introduces a NOTA loss function to improve the ability to identify NOTA relation (Section 3.4).

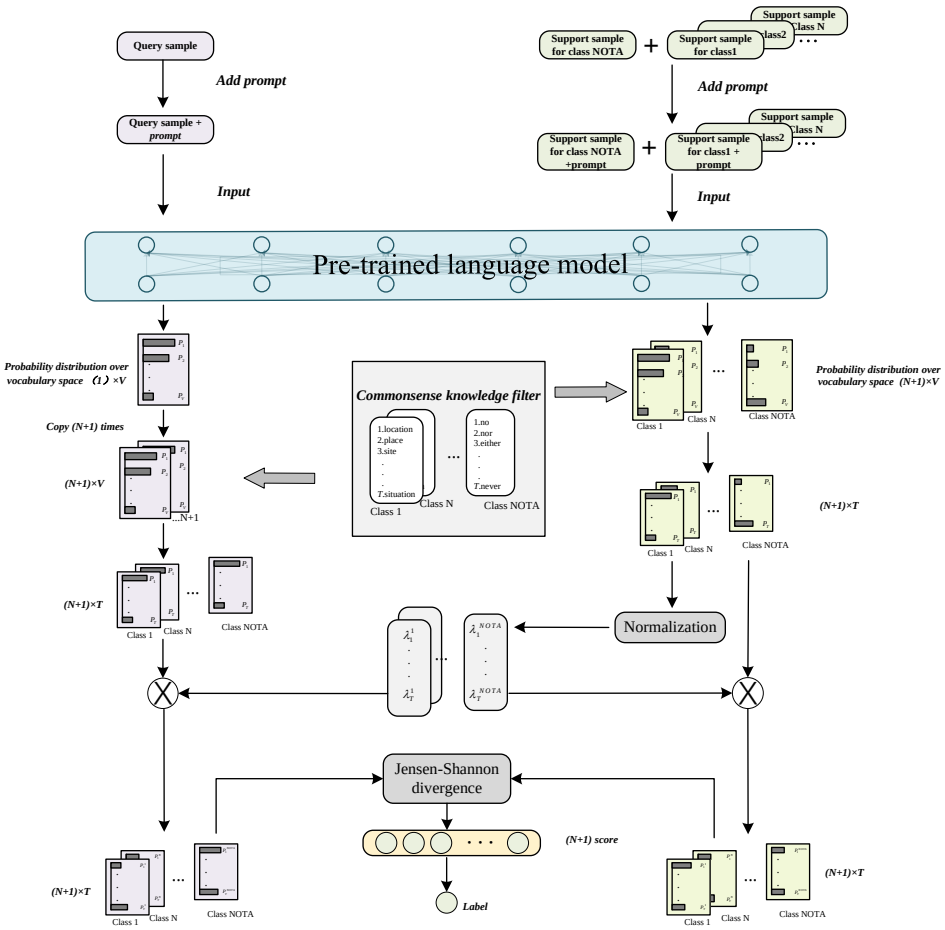


Figure 3. The framework of commonsense knowledge-aware prompt tuning. The right part shows the process of encoding input class supports, and the left part shows the process of encoding the input query.

3.1. Problem Definition

Given a training set $\mathcal{D}_{base}$ containing samples of base classes $\mathcal{C}_{base}$, the goal of the N-way few-shot NOTA relation classification task is to train a model with $\mathcal{D}_{base}$ to predict the relation r_q between the entity pair (h_q, t_q) mentioned in the query sentence q, where $r_q \in \mathcal{C}_{novel} = \{r_1, r_2, \ldots, r_N, NOTA\}$, and $\mathcal{C}_{novel} \cap \mathcal{C}_{base} = \emptyset$. That is, the novel classes are totally different from the base classes, and the number of class labels in the novel classes is $N+1$. In addition, for a K-shot task, each class label $r_i \in \mathcal{C}_{novel}$ is provided by K support samples, $S_i = \{s_{ij}\}_{j=1}^{K}$, where s_{ij} represents the j-th support sample sentence for class r_i, and s_{ij} contains two entities h_{ij} and t_{ij}. In summary, the task is to predict the label r_q of a query q according to the set of support samples, $\{S_i\}_{i=1}^{N+1}$.

3.2. Prompt Tuning Construction

Prompt tuning formalizes the classification task into a masked language modeling problem. Specifically, prompt-tuning wraps the input sentence with a natural language template, where several words are obscured that imply a relation between two entities contained in the sentence. For example, as show in Figure 2, an input query is "*London is the biggest city in the UK*". The entities contained in the query are kept. The converted input to the model is "*[CLS] + original sentence + In this sentence, the London is the [MASK] of the UK.[SEP]*". Similarly, a support sample for the relation class of the above query, "*Washington is located in the United States.*", is converted to the sentence, "*[CLS] +$x_{support}$+ In this sentence, the Washington is the [MASK] of the United States.[SEP]*". Let $\mathcal{M}$ be a language model pre-trained on large scale corpora (in this paper, we use BERT). Let $q = (x_1^q, h_q, x_3^q, \ldots, t_q, \ldots, x_n^q)$ be a query sentence, where h_q and t_q are two entities, and n is the length of the query. After preprocessing, the converted sentence of q is input to the $\mathcal{M}$. Its contextualized representation is produced, such as $\{h_{[CLS]}^q, h_1^q, \ldots, h_{[MASK]}^q, \ldots, h_{[SEP]}^q\}$. Finally, the representation of *[MASK]* is fed into an output layer of $\mathcal{M}$ to predict the probability distribution over the vocabulary space:

$$P_q(h_{[MASK]}^q = v, v \in V|q) = softmax(\mathbf{W}h_{[MASK]}^q + \mathbf{b}), \tag{1}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable model parameters, V is the vocabulary of the model, and v is a word in the vocabulary V. The above formula gets the probability that the relation contained in the question is v. Similarly, input the support samples of each class separately, and the similarity between each class and each word in the vocabulary is obtained. Let $s_{ij} = (x_1^{s_{ij}}, h_{ij}, x_3^{s_{ij}}, \ldots, t_{ij}, \ldots, x_m^{s_{ij}})$ be the j-th support sample for relation class r_i. Wrap s_{ij}, input to the pre-trained language model $\mathcal{M}$, and obtain the representation of *[MASK]*, $h_{[MASK]}^{s_{ij}}$. The similarity between the relation represented by this support sample, and each word in the vocabulary is

$$P_{ij}(h_{[MASK]}^{s_{ij}} = v, v \in V|s_{ij}) = softmax(\mathbf{W}h_{[MASK]}^{s_{ij}} + \mathbf{b}). \tag{2}$$

For the K-shot task, there are K support samples for each relation class. Thus, input K support samples separately and obtain K probability distributions. Finally, the mean-pooling operation is performed to integrate the probability distributions over the vocabulary space obtained by K support samples of a relation label:

$$P_{r_i}(r_i = v, v \in V|S_i) = POOL(P_{i1}(h_{[MASK]}^{s_{i1}} = v, v \in V|s_{i1}), \ldots, \tag{3}$$

$$P_{iK}(h_{[MASK]}^{s_{iK}} = v, v \in V|s_{iK})) \tag{4}$$

where $POOL$ represents a mean-pooling operation.

3.3. Commonsense Knowledge Enhanced Prompt-Tuning

The process of predicting masked words based on the context is not a single-choice procedure, that is, there is no standard correct answer. There are a wealth of words that may be suitable for this context. To reduce the uncertainty of predictions of the masked language model, we expand the label set. The expanded labels come from the synset of the original labels and form a new label space, which is screened from a commonsense knowledge base. Next, the label words that are split into multiple tokens are removed since they tend to be more tricky to handle in the training objective. Then, T similar extended labels are kept. Let V_i denote the subset of V, which is mapped into a specific label r_i. The learned weights are introduced to measure similarity between the input query q and the relation label r_i by the following function:

$$\lambda_k^i = \frac{exp(P_{r_i}(r_i = v_k^i, v_k^i \in V_i|S_i))}{\sum_{k=1}^{T} exp(P_{r_i}(r_i = v_k^i, v_k^i \in V_i|S_i))}. \tag{5}$$

where T is the number of extended labels for one original label. The similarity between the query q and the vocab v_k^i for class r_i is calculated by

$$P(v_k^i|q) = \lambda_k^i P_q(h_{[MASK]}^q = v_k^i, v_k^i \in V_i|q). \tag{6}$$

The probability distribution that the query q contains the extended label for relation r_i is

$$P_q^i = [P(v_1^i|q), P(v_2^i|q), \ldots, P(v_T^i|q)] \tag{7}$$

Similarly, the similarity between the class r_i and the extended label v_k^i for class r_i is calculated as

$$P(v_k^i|S_i) = \lambda_k^i P_{r_i}(r_i = v_k^i, v_k^i \in V_i|S_i). \tag{8}$$

The similarity between the original label r_i and extended labels for r_i is:

$$P_{r_i} = [P(v_1^i|S_i), P(v_2^i|S_i), \ldots, P(v_T^i|S_i)] \tag{9}$$

Since the impact of the label and query sample should be measured at the distribution level, we choose Jensen–Shannon divergence as a metric to measure the similarity of the two distributions. The more similar the samples, the smaller the value of the divergence distribution. Let $JS(\cdot||\cdot)$ denote the Jensen–Shannon divergence function. Thus, $1 - JS(\cdot||\cdot)$ is used to represent the similarity between the label and query sample. The probability score of the query q containing relation r_i is computed by

$$s(q, r_i) = T - JS(P_q^i||P_{r_i}) \qquad JS(\cdot||\cdot) \in [0, T]. \tag{10}$$

Specifically, the Jensen–Shannon divergence is calculated by

$$KL(Q(x)||P(x)) = \sum Q(x) log \frac{Q(x)}{P(x)}, \tag{11}$$

$$JS(Q(x)||P(x)) = \frac{1}{2} KL(P(x)||\frac{P(x) + Q(x)}{2}) + \frac{1}{2} KL(Q(x)||\frac{P(x) + Q(x)}{2}) \tag{12}$$

where $Q(x)$ and $P(x)$ are any two probability distributions.

3.4. A Learned Scoring Strategy

There are two situations in which the query should be classified as the NOTA class. One situation is that the two entities in the query do not have any relationship, and the other situation is that the two entities in the query have a certain relationship, but this relationship does not belong to any labels provided by the dataset. We design a learnable classification strategy to distinguish the NOTA relation by introducing a threshold parameter. The

threshold is determined by the labeled samples in the training set. Calculate the similarity between any two classes in the support set, and average them as the threshold

$$\theta = \frac{1}{(L-1)!}\sum_{i}^{L}\sum_{j=i+1}^{L} s(r_i, r_j), \tag{13}$$

where L represents the total number of classes in a support set. After obtaining the similarity between the query and each class, $\{s(q, r_i), r_i \in C_{novel}\}$, the classification strategy of the model is as follows. If the similarity score of the query and any class is lower than the threshold θ, the query belongs to the NOTA relation. Otherwise, the query belongs to the class with the highest similarity score.

When the relation contained in the query and class are negative pairs, their similarity is very low. After passing the softmax function, the probability of the similarity of negative sample pairs will be 0, which makes standard softmax prediction probability fail in the few-shot setting and the original cross-entropy formulation unsuitable for NOTA detection. Thus, a special term is added to the overall loss to accurately detect NOTA examples, in addition to accurately performing the episode's classification task. Intuitively, adding this term will change the embedding when training the model, making the optimized score perform well on the NOTA task. The overall loss function is as follows:

$$Loss(\{s\}) = -\sum_{r_{(target)} \in r^+} \log s(q, r_{(target)}) - \sum_{r_{(\neg target)} \in r^-} \tau \log(T - s(q, r_{(\neg target)})) \tag{14}$$

where r^+ is the positive sample set, r^- is the negative sample set, $\{s\}$ is the set of similarity scores between the query and each class, including the scores of the query and the positive class $\{s(q, r_{(target)}), r_{(target)} \in r^+\}$, and the scores of the query and the negative class $\{s(q, r_{(\neg target)}), r_{(\neg target)} \in r^-\}$, and τ is a penalty parameter.

4. Results

In this section, we conduct experiments to evaluate the effectiveness of our methods.

4.1. Dataset

As shown in Table 1, we use two fine-grained few-shot relation extraction datasets: FewRel 2.0 [1] and Few-Shot TACRED [39].

Table 1. Number of relation instances in the FewRel 2.0 and Few-Shot TACRED datasets.

Dataset	Train	Val	Test
FewRel 2.0	70,000	2500	3000
Few-Shot TACRED	8163	633	804

FewRel 2.0 [1] is a more challenging few-shot relation classification task with the none-of-the-above setting based on the N-way K-shot setting. It adopts the original FewRel training set for training and the newly annotated dataset for testing. Additionally, FewRel 2.0 includes the SemEval-2010 task 8 dataset as the validation set.

The Few-Shot TACRED [39] dataset was collected from a news corpus, purposing extracting relations involving 100 target entities. Accordingly, each sentence containing a mention of one of these target entities was used to generate candidate relation instances for the RC task. The relation label was annotated as 1 of 41 pre-defined relation categories, when appropriate, or into an additional *no* relation category. The *no* relation category corresponds to cases where some other relation type holds between the two arguments, as well as cases in which no relation holds between them. The Few-Shot TACRED dataset has a test set including 10 relations, a val set including 6 relations, and a training set including 25 relations.

4.2. Experimental Setup

The NOTA few-shot relation classification is based on the N-way K-shot setting. For the original N-way K-shot setting, each episode has a query instance q, the correct relation label $r_q \in \{r_1, r_2, \ldots, r_N\}$, and each class has K samples. For the NOTA few-shot classification, the correct relation label becomes $r_q \in \{r_1, r_2, \ldots, r_N, NOTA\}$ rather than $r_q \in \{r_1, r_2, \ldots, r_i\}$.

The parameter NOTA rate is to describe the proportion of NOTA queries at the test stage. For example, a 0% NOTA rate means no queries belong to the NOTA relation and the 50% NOTA rate means half of the queries have the NOTA label.

Accuracy is adopted as the evaluation metric

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

where $TP + TN$ is the number of queries correctly classified, and $TP + FP + FN + TN$ are the number of all queries.

4.3. Experimental Details

We use the BERT base [2] as the backbone structure of our model, initialized with the corresponding pre-trained cased weights. The hidden size is 768, and the number of layers is 12. Models are implemented by the Pytorch framework and Huggingface transformers. BERT models are optimized by AdamW with the learning rate of 6×10^{-5}. The training batch size used is 16 for all models. In the supervised setting, each model is trained for 10 epochs and evaluated on the dev set every 1000 steps. In the few-shot setting, each model is trained for 16 epochs and evaluated every 10–50 steps; each time the evaluation is run for 200 steps. Experiments are conducted with CUDA on NVIDIA Tesla V100 GPUs.

4.4. Models

CKPT—Our commonsense knowledge prompt tuning approach adding a commonsense knowledge expanded label on the basis of the prompt tuning approach.

Sentence Pair [1]—A fine-tuned BERT-based model utilizing the embedding-based next sentence prediction score of BERT [2] as the similarity score between a query and each support set instance.

Threshold [1]—A fine-tuned BERT-based model setting a predetermined threshold for NOTA few-shot classification. When the NOTA option is present, the NOTA class tests queries whose similarity with all of the target classes does not surpass the predetermined threshold.

NAV [3]—A fine-tuned BERT-based model for few-shot classification with the NOTA class. In this approach, the NOTA class is represented by an explicit vector in the embedding space, which is learned during training. At test time, the similarity between the query and this vector is computed and regarded as the probability that the query belongs to the NOTA relation.

MNAV [3]—A natural extension of the NAV approach, which is to represent the NOTA class by multiple vectors, whose value is an empirically tuned hyperparameter.

4.5. FewRel 2.0 Result

We first confirm the appropriateness of our investigation by comparing the performance of the prior FewRel 2.0 test data. Table 2 presents the figures on the two official (synthetic) test NOTA rates for this benchmark. We use the 50% NOTA rate to train all our models, with 3000 episodes per epoch. As shown, the CKPT model performs best across all FewRel settings, obtaining a new state of the art for this task.

We next turn to a comparison of the investigated embedding-based few-shot models on the FewRel 2.0 val set, with a 50% NOTA rate. Most of the previous models used a 50% NOTA rate for experiments. We chose a 50% NOTA rate to compare CKPT with previous models more comprehensively. The results in Table 2 show that the CKPT model

outperforms others in both settings. The gap between CKPT and the previous state-of-the-art model is significant for the two settings.

Overall, the prompt-tuning methods have shown certain improvements, compared to directly fine-tuned models. It shows that the prompt-based method does help with capturing contextual information from a given sentence. It is also observed that the magnitude of the improvement and the preference of the prompt encoding strategy may vary with different datasets. The prompt-tuning method seems less effective on the FewRel 2.0 test dataset than on the FewRel 2.0 val dataset. It indicates that the effect of the prompt-based method partially depends on the characteristics of the dataset and that different prompt designs may suit different data.

Table 2. Accuracy(%) results on FewRel 2.0 dataset for the four available settings for this benchmark.

Dataset	Model	1-Shot(15%)	1-Shot (50%)	5-Shot(15%)	5-Shot (50%)
FewRel 2.0 test	Sentence-Pair	77.67%	80.31%	84.19%	86.06%
	Threshold	63.41%	76.48%.	65.43 %	78.95%
	NAV	77.17%	81.47%	82.97%	87.08%
	MNAV	79.06%	81.69%	85.52%	87.74%
	CKPT	**80.37%**	**83.02%**	**86.26%**	**88.12%**
FewRel 2.0 val	Sentence-Pair	70.32%	75.48%	74.27%	78.43%
	Threshold	63.28%	76.32%	66.89%	80.30%
	NAV	-	78.54%	-	80.44%
	MNAV	-	78.23%	-	81.25%
	CKPT	**73.28%**	**81.25%**	**77.92%**	**83.62%**

4.6. Few-Shot TACRED Results

We compare the CKPT, MNAV, NAV, sentence-pair and threshold-based models over the Few-Shot TACRED test set. As seen in Table 3, the performance of CKPT is better than others, just like the situation for FewRel 2.0. Due to several differences between the datasets, including training size, NOTA rate, and different entity types, the results on Few-Shot TACRED are drastically lower than those obtained for FewRel 2.0. The most important reason is that the training data on FewRel TACRED are significantly less than the training data on FewRel 2.0, and the ability of the traditional fine-tune method to train the model to learn to measure the gap between samples is significantly weakened. In contrast, the prompt tuning method converts downstream tasks into tasks similar to the pre-train stage to fully tap the potential of the pre-training model. In addition, the fewer the labeled data, the greater the alignment with the theme of small sample learning. The pre-training model uses the self-supervised masked language mechanism to learn a large number of text features in the pre-train stage and uses the prompt tuning method to apply this part of the information in the classification to reduce the dependency of downstream tasks on annotated data.

Table 3. Micro F1 results on Few-Shot TACRED.

Model	5-Way 1-Shot	5-Way 5-Shot
Sentence-Pair	10.19 ± 0.81%	-
Threshold	6.87 ± 0.48%	13.57 ± 0.46%
NAV	8.38 ± 0.80%	18.38 ± 2.01%
MNAV	12.39 ± 1.01%	30.04 ± 1.92%
CKPT	**15.14 ± 1.12%**	**32.26 ± 2.13%**

4.7. Ablation Experiments

In this section, we conduct ablation studies to analyze how each component affects the few-shot recognition performance. Table 4 shows the results of our ablation studies on the FewRel 2.0 development set. PT is a prompt tuning approach based on BERT [2]

and uses the traditional cross entropy loss function. NOTA-Loss is a novel loss function approach Section 3.4 for few-shot NOTA classification. PT+NOTA-Loss is a prompt tuning approach using the novel loss function approach (Section 3.4) and does not use the method of extending the label. It indicates that the model using NOTA-Loss achieves better results compared with the model using the traditional cross-entropy loss. This is because NOTA-Loss uses coefficients to amplify the loss of two samples that are not related to the two entities, and the model learns more about these parts of the samples. In addition, the label of the masked language model in the pre-train stage is the entire vocab text. Therefore, in the prompt tuning prediction stage, the model may predict words with similar meanings to the label, causing model classification errors. As shown in Table 4, we use commonsense knowledge to expand the label, which can reduce the contingency of the masked language model for label prediction and improve the prediction performance of the model.

Table 4. An ablation study of our proposed method on the FewRel 2.0 development set.

Model	5-Way 1-Shot (50%)	5-Way 5-Shot (50%)
PT	79.64 ± 0.10%	82.25 ± 0.13%
PT + NOTA-Loss	80.35 ± 0.15%	82.76 ± 0.14%
CKPT	**81.25 ± 0.12%**	**83.62 ± 0.13%**

4.8. Effect of Templates

The choice of templates may have a huge impact on the performance in prompt learning. In this section, we carry out experiments to investigate such an influence. The results demonstrate that the choice of templates exerts a considerable influence on the performance of prompt-based few-shot learning. As shown in Table 5, the phrase that describes the location "in this sentence" contributes a remarkable improvement in performance. Specifically, as we only change the direction of the relations and yield such improvements, prompts are position aware. Therefore, the automatic selection of different templates for different datasets is also the main direction of our future research.

Table 5. Effect of templates. The results are produced under the development set dataset by CKPT.

Template	5-Way	
	1-Shot (50%)	5-Shot (50%)
x.the *entity*1 is the [MASK] the *entity*2.	80.34 ± 0.11%	82.16 ± 0.18%
x.In this sentence, the *entity*2 is the [MASK] the *entity*1.	80.95 ± 0.12%	82.87 ± 0.14%
x.In this sentence, the *entity*1 is the [MASK] the *entity*2.	**81.25 ± 0.12%**	**83.62 ± 0.13%**

4.9. NOTA Rates Impact

We control the unrealistic NOTA rate in FewRel 2.0 by training and evaluating our model on higher NOTA rates. The results in Figure 4 indicate that as the NOTA rate increases, the rate of decrease in the accuracy of CKPT is significantly less than that of MNAV. This is because pre-trained language models learn a lot of predictive capabilities related to NOTA in the unsupervised pre-training stage. Compared with the pre-trained model, CKPT can predict the NOTA relation more accurately.

At the same time, it can be observed that the predicted F1 value of the model will become worse as the NOTA rate increases, mainly because the model is difficult to judge the None category without any relationship between entities. The "None" category is not only a relation that has never appeared in the training set, but also has many texts from different corpora. There will be some differences in different corpora, which undoubtedly increases the difficulty of model judgment. Therefore, improving the robustness of the pre-trained model is also our main work in the future.

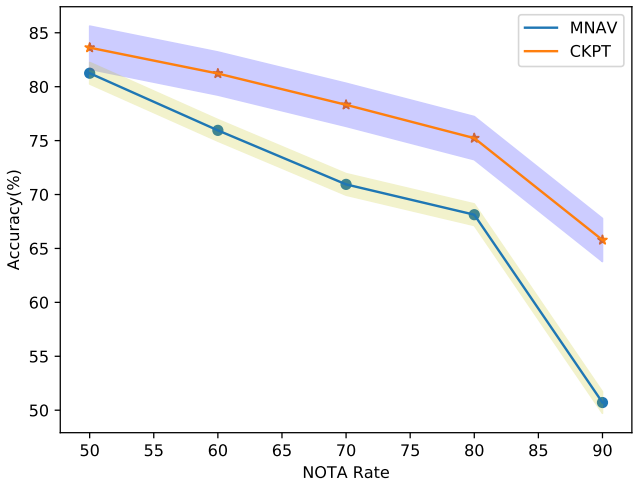
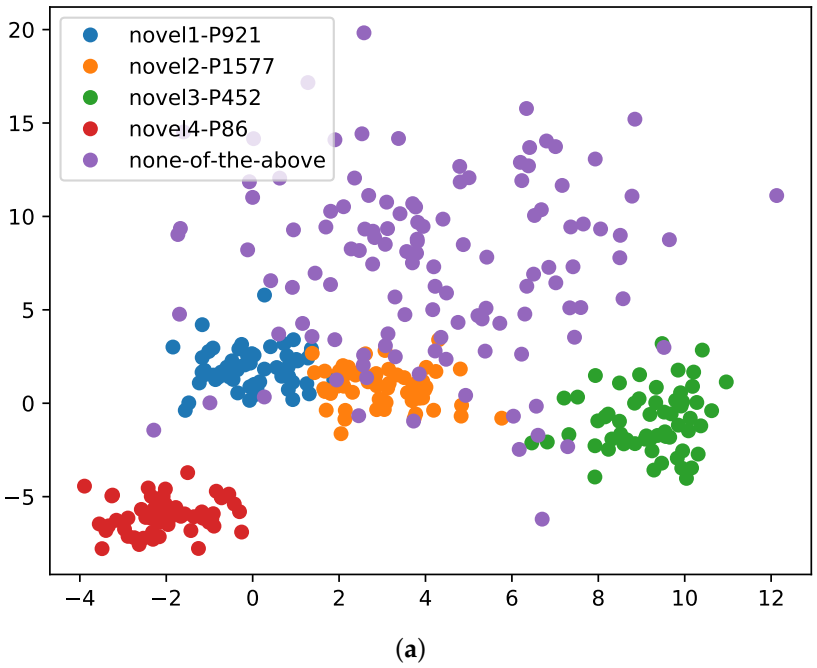


Figure 4. CKPT and MNAV results on the FewRel 2.0 dev dataset at different NOTA rates.

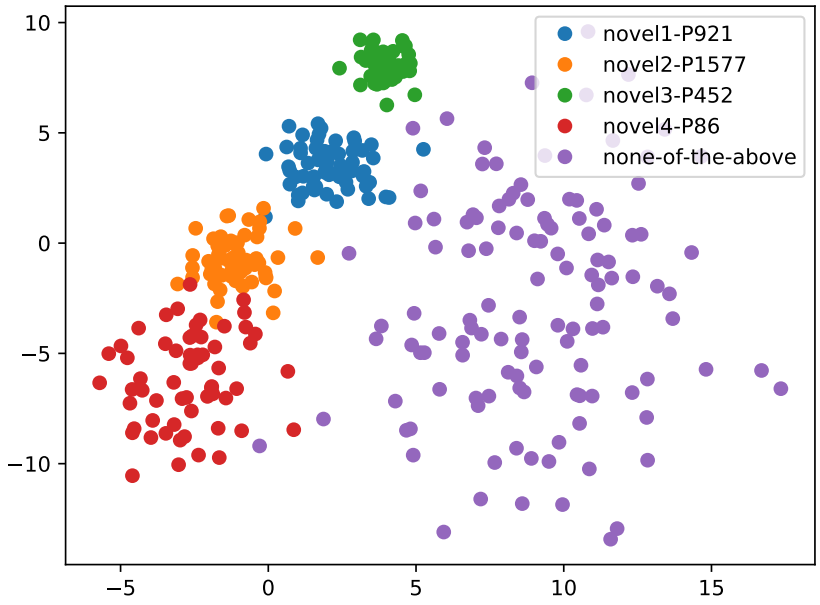
4.10. Effect of NOTA Loss

Figure 5 provides the visualization of the t-SNE-transformed feature representations. We can observe that for the model using softmax cross-entropy loss, some features of positive samples and negative samples are mixed, and the boundary between positive and negative samples is not clear. Traditional softmax cross-entropy loss causes the loss of negative sample pairs to be too small, resulting in insufficient model learning and the overlap of the distribution of positive and negative samples in the embedding space. To solve this problem, NOTA loss is introduced to learn the uniform distribution of negative classes in the feature embedding space by amplifying the negative sample loss and using the coefficient factor to control the proportion of negative sample loss and positive sample loss. It can be seen from the Figure 5 that NOTA loss enables the model to fully learn the characteristics of negative samples and improves the distribution of positive and negative sample pairs in the embedding space.



(**a**)

Figure 5. *Cont.*



(**b**)

Figure 5. A t-SNE plot of the computed feature representations of instances in the FewRel 2.0 val set. Node colors denote relation classes: P921 is a number of relationship classes, and the none-of-the-above denote the NOTA relation. (**a**) Softmax cross entropy loss; (**b**) NOTA loss.

4.11. Effect of the Number of Expanding Label Words

Expanding the number of labels has an important impact on the experimental results. In order to find a suitable number of labels, we have conducted many experiments. As shown in Figure 6, when the number of extended labels is small, increasing the number of labels can greatly improve the results of the experiment, but when the number of labels exceeds 50, the accuracy of the experiment gradually decreases. This is because the number of highly relevant synonyms in the knowledge base is limited. As the number of extended labels increases, the correlation between the extended labels and the source labels becomes lower and lower. These low-correlation extended labels become noise that affects the final experimental accuracy. Therefore, in this article, all experiments use an extended label number of 50.

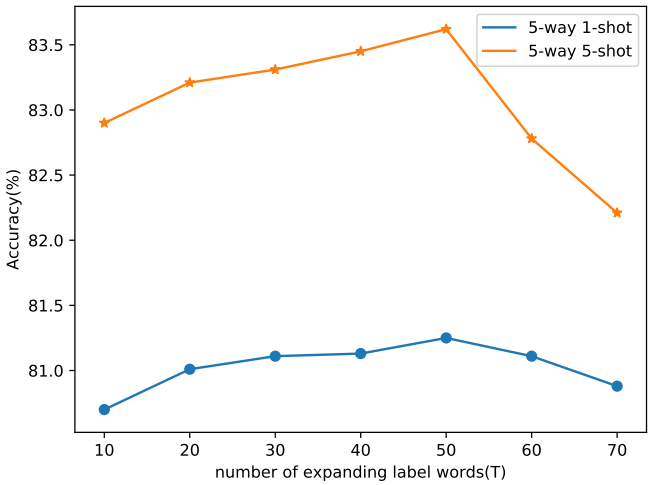


Figure 6. Effect of the number of expanding label words, and the NOTA rate used in the experiment is 50%.

4.12. Run-Time of the Model Prediction

As shown in Table 6, we conducted experiments on the prediction time of the model on the FewRel 2.0 dataset. The predictions used in the experiments are 5way-1shot and 5way-5shot, respectively. It can be seen from the figure that the prediction time of CKPT is the shortest, because the model structure of CKPT is simpler, and the parameters of the model are smaller than those of NAV and MNAV. At the same time, the current pre-training model has a large number of parameters, and the overall inference speed is still very slow. In the future, we will also study knowledge distillation, quantization and other methods to build a pre-training model for fast inference of our tasks.

Table 6. The run-time of the model prediction.

Model	5-Way 1-Shot	5-Way 5-Shot
NAV	0.52 s	2.62 s
MNAV	0.68 s	3.12 s
CKPT	**0.47 s**	**2.28 s**

5. Conclusions

In this paper, we contribute to the few-shot NOTA relation classification with a concise and effective prompt tuning baseline named commonsense knowledge-aware prompt tuning. We propose a commonsense knowledge-enhanced method for prompt tuning that injects commonsense knowledge into the prompt label construction in order to express the NOTA relation more comprehensively. We design a learned scoring strategy on top of the embedding of our model, which is specially designed for the NOTA task combined with the prompt-tuning method to more accurately identify the NOTA class. Experiments show that our method achieves a new state of the art in the field of few-shot NOTA classification, indicating that the use of the prompt tuning method to classify samples is a promising direction for future research.

Our approach can also be applied in areas such as cultural heritage [40] and labor market analysis [41]. However, commonsense knowledge-aware prompt tuning methods are handcrafted and somewhat straightforward. A natural direction for improving it is training an additional convolutional neural network end to end to measure the transductive similarity.

In our experiments, we found that different templates have a great impact on the accuracy of the pre-training model, and the training method of the pre-training model during the pre-training process also has a great impact on the experimental results. Therefore, there are two important directions for future work: (1) design a unified task format and corresponding pre-training objectives for other types of tasks, such as language generation and relation extraction, and (2) build an automatic template generation tool to generate different templates for different tasks.

Author Contributions: Conceptualization, B.L. and L.J.; methodology, B.L. and L.J.; validation, B.L., Y.Z. and H.W.; formal analysis, B.L. and X.L.; investigation, B.L.; resources, H.W.; data curation, L.J.; writing, B.L.; visualization, Y.Z.; funding acquisition, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences grant no. Y835120378.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare that they do not have any conflict of interest. This research does not involve any human or animal participation. All authors have checked and agreed with the submission.

Abbreviations

The following abbreviations are used in this manuscript:

NOTA	None-of-the-above
PLMs	Pre-trained language models
CKPT	Commonsense knowledge-aware prompt tuning

References

1. Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. 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), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 6250–6255.
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*; NAACL: Baltimore, MD, USA, 2019.
3. Sabo, O.M.S.; Elazar, Y.; Goldberg, Y.; Dagan, I. Revisiting Few-shot Relation Classification: Evaluation Data and Classification Schemes. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 691–706. [CrossRef]
4. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 16 January 2022).
5. Sampath, P.; Sridhar, N.S.; Shanmuganathan, V.; Lee, Y. TREASURE: Text Mining Algorithm Based on Affinity Analysis and Set Intersection to Find the Action of Tuberculosis Drugs against Other Pathogens. *Appl. Sci.* **2021**, *11*, 6834. [CrossRef]
6. Zhang, H.; Zhang, G.; Ma, Y. Syntax-Informed Self-Attention Network for Span-Based Joint Entity and Relation Extraction. *Appl. Sci.* **2021**, *11*, 1480. [CrossRef]
7. Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; Sun, M. 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, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 4803–4809.
8. Dou, Z.Y.; Yu, K.; Anastasopoulos, A. Investigating Meta-Learning Algorithms for Low-Resource Natural Language Understanding Tasks. 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), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 1192–1197.
9. Satorras, V.G.; Estrach, J.B. Few-Shot Learning with Graph Neural Networks. In Proceedings of the 2018 International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
10. Mishra, N.; Rohaninejad, M.; Chen, X.; Abbeel, P. A Simple Neural Attentive Meta-Learner. In Proceedings of the 2018 International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
11. Bengio, Y.; Bengio, S.; Cloutier, J. Learning a synaptic learning rule. In Proceedings of the IJCNN-91-Seattle International Joint Conference on Neural Networks, Seattle, WA, USA, 8–12 July 1991; Volume II, p. 969. [CrossRef]
12. Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. In Proceedings of the 2017 International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
13. Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. *arXiv* **2017**, arXiv:1707.09835 .
14. Geng, X.; Chen, X.; Zhu, K.Q.; Shen, L.; Zhao, Y. MICK: A Meta-Learning Framework for Few-shot Relation Classification with Small Training Data. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020.
15. Wang, Y.; Bao, J.; Liu, G.; Wu, Y.; He, X.; Zhou, B.; Zhao, T. Learning to Decouple Relations: Few-Shot Relation Classification with Entity-Guided Attention and Confusion-Aware Training. In Proceedings of the 28th International Conference on Computational Linguistics, Online, 8–13 December 2020; International Committee on Computational Linguistics: Barcelona, Spain, 2020; pp. 5799–5809.
16. Dong, B.; Yao, Y.; Xie, R.; Gao, T.; Han, X.; Liu, Z.; Lin, F.; Lin, L. Meta-Information Guided Meta-Learning for Few-Shot Relation Classification. In Proceedings of the 28th International Conference on Computational Linguistics, Online, 8–13 December 2020; International Committee on Computational Linguistics: Barcelona, Spain, 2019; pp. 1594–1605.
17. Qu, M.; Gao, T.; Xhonneux, L.P.; Tang, J. Few-shot Relation Extraction via Bayesian Meta-learning on Relation Graphs. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; Volume 119, pp. 7867–7876.
18. Seo, C.W.; Seo, Y. Seg2pix: Few Shot Training Line Art Colorization with Segmented Image Data. *Appl. Sci.* **2021**, *11*, 1464. [CrossRef]
19. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208. [CrossRef]

20. Ren, H.; Cai, Y.; Chen, X.; Wang, G.; Li, Q. A Two-phase Prototypical Network Model for Incremental Few-shot Relation Classification. In Proceedings of the 28th International Conference on Computational Linguistics, Online, 8–13 December 2020; International Committee on Computational Linguistics: Barcelona, Spain, 2020; pp. 1618–1629. [CrossRef]
21. Fan, M.; Bai, Y.; Sun, M.; Li, P. Large Margin Prototypical Network for Few-shot Relation Classification with Fine-grained Features. In Proceedings of the 2019 ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; Zhu, W., Tao, D., Cheng, X., Cui, P., Rundensteiner, E.A., Carmel, D., He, Q., Yu, J.X., Eds.; ACM: New York, NY, USA, 2019; pp. 2353–2356.
22. Ding, N.; Wang, X.; Fu, Y.; Xu, G.; Wang, R.; Xie, P.; Shen, Y.; Huang, F.; Zheng, H.T.; Zhang, R. Prototypical Representation Learning for Relation Extraction. In Proceedings of the 2021 International Conference on Learning Representations, Vienna, Austria, 30 April–3 May 2021.
23. Wu, L.; Zhang, H.P.; Yang, Y.; Liu, X.; Gao, K. Dynamic Prototype Selection by Fusing Attention Mechanism for Few-Shot Relation Classification. In *ACIIDS (1)*; Nguyen, N.T., Jearanaitanakij, K., Selamat, A., Trawinski, B., Chittayasothorn, S., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2020; pp. 431–441.
24. Tan, M.; Yu, Y.; Wang, H.; Wang, D.; Potdar, S.; Chang, S.; Yu, M. Out-of-Domain Detection for Low-Resource Text Classification Tasks. 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), Hong Kong, China, 3–7 November 2019 ; Linguistics: Hong Kong, China, 2019; pp. 3566–3572.
25. Teigen, A.L.; Saad, A.; Stahl, A.; Mester, R. Few-Shot Open World Learner. *IFAC-PapersOnLine* **2021**, *54*, 444–449. [CrossRef]
26. Willes, J.; Harrison, J.; Harakeh, A.; Finn, C.; Pavone, M.; Waslander, S.L. Bayesian Embeddings for Few-Shot Open World Recognition. *CoRR* **2021**. Available online: http://xxx.lanl.gov/abs/2107.13682 (accessed on 16 January 2022) .
27. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
28. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *CoRR* **2021**. Available online: http://xxx.lanl.gov/abs/2107.13586 (accessed on 16 January 2022).
29. Schick, T.; Schütze, H. 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, Online, 10 September 2020.
30. Han, X.; Zhao, W.; Ding, N.; Liu, Z.; Sun, M. PTR: Prompt Tuning with Rules for Text Classification. *arXiv* **2021**, arXiv:2105.11259.
31. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
32. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the NAACL-HLT 2019: Demonstrations, Minneapolis, MN, USA, 2–7 June 2019.
33. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
34. Ding, N.; Chen, Y.; Han, X.; Xu, G.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; Kim, H. Prompt-Learning for Fine-Grained Entity Typing. *CoRR* **2021**. Available online: http://xxx.lanl.gov/abs/2108.10604 (accessed on 16 January 2022).
35. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3816–3830. [CrossRef]
36. Shin, T.; Razeghi, Y.; Logan, R.L., IV; Wallace, E.; Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16-20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4222–4235. [CrossRef]
37. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021.
38. Hambardzumyan, K.; Khachatrian, H.; May, J. WARP: Word-level Adversarial ReProgramming. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4921–4933. [CrossRef]
39. Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; Manning, C.D. Position-aware Attention and Supervised Data Improve Slot Filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 35–45.

40. Amato, F.; Moscato, F.; Moscato, V.; Pascale, F.; Picariello, A. An agent-based approach for recommending cultural tours. *Pattern Recognit. Lett.* **2020**, *131*, 341–347. [CrossRef]
41. Colace, F.; De Santo, M.; Lombardi, M.; Mercorio, F.; Mezzanzanica, M.; Pascale, F. Towards Labour Market Intelligence through Topic Modelling. In Proceedings of the Hawaii International Conference on System Sciences, Maui, HI, USA, 8–11 January 2019. [CrossRef]