Preprint submitted to Computing and Informatics

ABILITIES OF CONTRASTIVE SOFT PROMPTING FOR OPEN DOMAIN RHETORICAL QUESTION DETECTION

Josef Baloun, Jiří Martínek

Dept. of Computer Science & Engineering
University of West Bohemia
Plzeň, Czech Republic
e-mail: {balounj, jimar}@kiv.zcu.cz

Cristophe Cerisara

CNRS LORIA,
Université de Lorraine
Nancy, France
e-mail: cerisara@loria.fr

Pavel Král

Dept. of Computer Science & Engineering
University of West Bohemia
Plzeň, Czech Republic
e-mail: pkral@kiv.zcu.cz

Abstract. In this work, we start by demonstrating experimentally that rhetorical question detection is still a challenging task, even for state-of-the-art Large Language Models (LLMs). We then propose an approach that boosts the performances of such LLMs by training a soft prompt in a way that enables building a joint embedding space from multiple loosely related corpora. The advantages of using a soft-prompt compared to finetuning is to limit the training costs and combat overfitting and forgetting. Soft prompting is often viewed as a way to guide the model towards a specific known task, or to introduce new knowledge into the model

through in-context learning. We further show that soft prompting may also be used to modify the geometry of the embedding space, so that the distance between embeddings becomes semantically relevant for a target task, similarly to what is commonly achieved with contrastive finetuning. We exploit this property to combat data scarcity for the task of rhetorical question detection by merging several datasets into a joint semantic embedding space. We finally show on the standard Switchboard dataset that the resulting BERT-based model nearly divides by 2 the number of errors as compared to Flan-T5-XXL with only 5 few-shot labeled samples, thanks to this joint embedding space. We have chosen in our experiments a BERT model because it has already been shown with S-BERT that contrastive finetuning of BERT leads to semantically meaningful representations. Therefore, we also show that this property of BERT nicely transfers to the soft-prompting paradigm. Finally, we qualitatively analyze the resulting embedding space and propose a few heuristic criteria to select appropriate related tasks for inclusion into the pool of training datasets.

Keywords: Soft Prompts, Prompt-tuning, Rhetorical Questions, Contrastive Learning, Triplet Loss, Pre-trained Language Models

1 INTRODUCTION

Nowadays, large pre-trained language models (PLM) are essential components of many natural language processing (NLP) applications. It has been demonstrated in many works that they encode valuable linguistic information, especially at the level of morphology, syntax, and semantics, as well as factual/world knowledge, such as cities, famous people, and historical events. It is also possible to extract procedural knowledge and multi-step reasoning from the largest PLMs (chain-of-thoughts, tree-of-thoughts...), for instance, to solve maths problems.

However, there are still gaps in some types of information that these PLMs hardly capture, or maybe we just do not know yet how to extract it from the PLM. Examples of such challenges are detecting implicatures and rhetorical questions (RQs). For such tasks, the majority of popular PLMs with prompt-based approaches fail, and PLM performance is close to random guessing, when considering zero-shot or few-shot scenarios.

This is shown for instance in [1], where the authors evaluate several PLMs with a prompt-based strategy for the task of conversational implicatures. The task is to automatically determine the implicature (i.e., the meaning yes or meaning no) for the question and its response (e.g., question: "Can you make a cake?", response: "Can birds fly?"[1]). Most models in their experiments obtained around 60% accuracy on the test set (random guessing performance is 50%) despite multiple prompt templates. Furthermore, the rhetorical question implicature (similar to the

[1] The response is a rhetorical question that semantically means the answer "yes".

example above) seems to be especially difficult to detect compared to the other implicature types (e.g., "world knowledge").

A natural option to try and solve these challenges would be to find datasets annotated with rhetorical questions and finetune a PLM on them. However, such datasets are relatively rare, small and heterogeneous, and finetuning large PLM incurs a prohibitive computational cost. We address the latter issue by exploiting soft-prompting ([2, 3, 4]) instead of finetuning, and show that contrastive training of the soft prompt enables to build a semantic embedding space adapted to the task, which, to the best of our knowledge, has never been demonstrated before. We further show that this embeddings space may be obtained from several related datasets, which are thus projected into a joint embedding space. We finally solve the former issue by adapting this embedding space to the target dataset in a few-shot way, without exploiting any development corpus at this stage.

1.1 Research question: building a joint and semantic embeddings space.

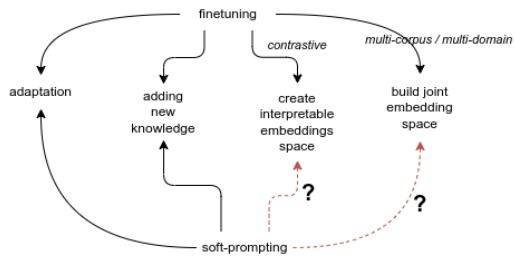

Fig. 1. Illustration of the research questions addressed in this work: does contrastive training of soft prompts enable to build a semantic embedding space for rhetorical questions? Does multi-corpus training of soft-prompting enable to build a joint embedding space across domains?

Finetuning the parameters of a LLM may be used to achieve two main purposes: either adapting the model to a new domain or style, or adding new knowledge. It has been shown previously that both objectives may also be achieved with parameter-efficient training, and soft-prompting in particular. However, while finetuning an LLM contrastively may be used to produce a semantic embeddings space, it is still unclear whether the same can be achived with soft prompting. The authors in [5] provide some hints that it might be possible, but with adapters and text-vision models. Nevertheless, such a conclusion is far from obvious given that soft prompting only affects the input sequence, which is a priori not designed to modify the global properties of the embeddings space; if this is true, as our next experiments suggest it is, then this might mean that it may be possible to alter the embeddings space just by manually crafting carefully designed prompts.

Similarly, it is well-known that joint embedding spaces across diverse domains or corpora may be obtained through multi-corpus finetuning, but whether parameter efficient training enables to do the same is still an open question.

We propose to experimentally demonstrate, on the challenging task of rhetorical question detection, that both objectives may also be reached with properly training the soft-prompt of a LLM.

1.2 Contributions

1. **Progress in rhetorical question detection:** We gather several datasets related to rhetorical question detection and propose a parameter-efficient approach that outperforms by a large margin the state-of-the-art (SOTA) open-source and reproducible Flan-T5-XXL.
2. **Claim and support two novel properties of soft-prompting:** We show that contrastive soft-prompting may also be used to create semantic embeddings space; and to build a joint multi-corpus embedding space.
3. **New insights about desirable structure of a joint embedding space for rhetorical questions:** We qualitatively analyse the resulting semantic embedding space and exhibit some interesting structure, e.g., the fact that the sarcastic feature is not well aligned with the binary rhetorical feature.
4. **Few-shot without development corpus:** when finetuning a system with few-shots, hyper-parameters must still be adjusted, which is often done on a development corpus. The fact to use a development corpus questions whether the model is really few-shot. We describe a methodology that does not require any development corpus and thus guarantees that the resulting model is indeed truly few-shot.

2 CORPORA

Rhetorical questions (RQs) typically occur in two data sources: dialogues and social networks – 42% of all questions on English Twitter are rhetorical [6]. In dialogues, though, rhetorical questions are among the least frequent categories of dialogue acts (see MRDA [7] and/or Switchboard [8] corpora). Therefore, methods working with little training data are required. Although there are some previous contributions for RQ detection in Twitter, e.g., [9], but the corpora change over time as tweets are deleted.

In this work, we focus on the automatic detection of RQs in dialogues. Indeed, detecting rhetorical questions is important for advanced conversational agents that target a more natural and smooth conversation. However, as summarized in the related works section, a variety of subtly different tasks exist around this notion of rhetorical questions. Therefore, we will study in this work corpora that contain rhetorical question labels (e.g., Switchboard, Meeting Recorder Dialog Act Cor-

pus – MRDA) as well as corpora dealing with "non-information seeking questions" (RQuet) and containing "sarcastic questions" (Sarcasm V2).

We included the following corpora in this work. All contain dialogues with RQs or related labels:

- The Switchboard Dialogue Act (**SWDA**) corpus is a set of manually transcribed conversations resulting in more than 220,000 utterances with 42 labels. We use the pre-processed balanced version of the SWDA corpus described in [10].
- The **MRDA** corpus contains over 100,000 hand-annotated dialogue act labels from dozens of naturally-occurring meetings. The corpus contains three levels of annotations; we use the RQ tags from the set of general labels;
- The **RQuet** corpus[2] [11] contains information seeking questions (ISQs) and non-information seeking questions (NISQs);
- The Sarcasm v2 (**SarcV2**) corpus[3] [12] contains more than 6,500 sarcastic and non-sarcastic internet posts. We consider this additional corpus because the task might be related to RQs and because we want to study the possibility to transfer information from a related but distant task.

The statistics are provided in Appendix A.

3 RELATED WORK

3.1 Prompt-based Zero-/Few-shot Approaches

Fine-tuning pre-trained language models is still a method of choice to adapt PLMs to a new task or domain, but the cost required to update every single parameter of the model might be cumbersome, especially with large PLM with hundreds of billions of parameters. A possible solution is proposed by [13] with text prompts strategy. The authors show that text prompts are very effective with a frozen GPT-3 model, and obtain zero-shot predictions just by using a natural language sentence before the actual text input, such as: "*Translate the following English sentence to German:* $<$ *sentence* $>$" $\rightarrow$ $<$ *germansentence* $>$. In-context few-shot learning may be achieved by showing a few task demonstrations (e.g., several examples of translations). Either way, there are no gradient updates and no fine-tuning. We rely on the model's ability to "understand" a prompt template.

However, such discrete (hard) prompts have several drawbacks. The main problem is that prompts that humans consider logical and suitable are not necessarily effective for language models. Besides, we do not know whether our prompt is optimal in the context of the domain and the model we use. Last but not least, the pre-trained models are sensitive to the choice of prompts [14]. The aforementioned

[2] `https://github.com/kkalouli/RQueT`

[3] `https://nlds.soe.ucsc.edu/sarcasm2`

problems have resulted in seeking "optimal" prompts in an automatical way relying on gradient-based searching [**?**] or automatic prompt generation [16].

Another set of methods is the usage of continuous (or soft) prompts [2, 17, 18]. The main idea behind this is adding additional learnable parameters (tunable "virtual text tokens") that are injected into a PLM. Usually, these tunable tokens are somehow appended to the input text and optimized while keeping the whole model frozen. This training paradigm allows us to share large-scale models for various tasks more efficiently (only a small portion of optimized parameters are stored). Approaches based on prompt tuning have shown interesting and promising results compared to the manually created prompts [2]. The soft-prompting strategy belongs to the more general family of parameter-efficient training methods, which also include for instance low-rank adaptation, adapters and prefix tuning.

3.2 Rhetorical Question Detection

In dialogues, question types may be clustered into two groups: factoid (information-seeking) and non-factoid [11]. The NLP community often focuses on factoid questions (e.g., in the Question Answering task), as answering questions such as "Who was the president of the USA before Clinton?" is common in NLP applications, while the second group of questions plays a more subtle but still important role in the course of dialogues, e.g., to emphasize a statement or implicitly criticize a position. The most significant type of such questions is rhetorical questions, even though several sub-types exist [19].

Primary attention in the past has been paid to rhetorical questions [20, 21, 22, 23, 24], mainly focusing on linguistic aspects. Automatic detection of RQs in Switchboard corpus has been presented in [10]. The authors used a set of features combined with a SVM. They obtained good results on a balanced version of the SWDA corpus, with 81% of accuracy without context, and 84% when a context is used.

Authors in [11] created the corpus RQuet – Resource of Question Types, a collection of information-seeking questions (ISQs) and non-information-seeking-questions (NISQs). The authors evaluated several features, including speaker change, prior and subsequent contexts, and part-of-speech tags. They obtained a recognition accuracy of around 77% when using questions without context.

Martinek et al. [25] focused on question type detection (including RQs) on the MRDA corpus. The authors used a linguistic rule-based approach together with a pre-trained BART [26] model to perform a zero-shot approach and found out that RQs are problematic for zero-shot.

Authors of [27, 12] have targeted the problem of rhetorical questions and sarcasm detection. As a source, they used debate forums, and labeling RQs was done automatically using heuristic rules.

In recent years, some efforts have been made to identify rhetorical questions in Twitter, notably [28, 29]. Zhuang et al. [9] created a corpus containing tweets with questions (rhetorical and information-seeking) and carried out experiments with

SVM and Bidirectional LSTM. The authors also integrated the prior tweet and a topic feature into the model and concluded that they seem to be helpful for this task.

Overall, the usage of Twitter corpora is problematic for reproducibility since a fraction of the tweets are regularly deleted over time, e.g., when a user deletes an account. We thus preferred to focus our work on dialogue corpora, such as MRDA and SWDA. Although RQs are less frequent in these corpora, long-term availability of data is guaranteed.

4 PRELIMINARY EXPERIMENTS

Before we delve into presenting our approach in detail, we describe the set of preliminary experiments demonstrating PLMs performance in the field of RQ detection. We focus on zero- or few-shot experiments with several open-source state-of-the-art PLMs. These experiments clearly show that all models dramatically fail at recognizing rhetorical questions, both in zero-shot (ZSL) and in-context few-shot (IC-FSL) setups. We explain this failure by the fact that recognizing rhetorical questions involves a reflexive, or meta-reasoning process, where the reader analyzes the semantic content of the question but also the intent of the speaker who is asking this question, intent that is different from the semantic content of the question.

Hence, the question "Do bicycles allow one to move from place to place?" might be interpreted by the PLM as a simple question with a clear semantic content, because the PLM's training corpus contains many direct questions that look similar at first view. But recognizing this question as a rhetorical one involves another level of reasoning about the context and the hidden intent of the speaker, who is likely not really interested in the factual answer to this question.

We show that such reflexive analysis are still out of reach of current open PLMs in Zero-Shot and IC-FSL modes, when 0 to 3 training samples per class are considered. We thus propose in Section 5 a solution that addresses the major flaws of the mainstream approaches commonly used to solve similar research questions, in particular fine-tuning and soft-prompt tuning. In these experiments, we test the following open state-of-the-art models:

- **Flan-T5-XXL** is a 11B-parameters model released by Google. It is considered as one of the best open PLM that outperforms GPT3 [30]. Because this model is an encoder-decoder transformer from the family of T5, we tested it with the following prompt templates and simply checked whether the answer is yes or no; note that it was always either "yes" or "no" in our experiments:
  - P1: "[input] Is the previous question a rhetorical question, yes or no?"
  - P2: "[input] Does the previous question require an answer, yes or no?"
  - P3: "Is the question a rhetorical question, yes or no? Question: [input]"
- **OPT-6.7B** is a 6.7B-parameters model released by Meta-AI [31]. It is one of the best open PLM from the GPT family. Since this model is a decoder-only

transformer from the family of GPT, the question-answering strategy used for Flan-T5-XXL does not work, as other words than yes or no were often generated. We then used the P1 and P2 prompts and chose the minimum perplexity when either "yes" or "no" was concatenated to the prompt.

- **TK-11b** is a 11B-parameters model released by Allen-AI [32]. It is an open PLM trained with the "instruction-tuning" strategy that has led, along with reinforcement learning from human feedback (RLHF), to the success of ChatGPT. We followed the recipe given with TK-11b. We used the following prompt, and selected the lowest perplexity among "yes" and "no" continuation because all answers are not yes/no:
  - P4: "Definition: is the following question a rhetorical question, yes or no? Input: [input] Output:"

In Table 1, we present the results of the zero-shot evaluation (see column 0), and the few-shot evaluation with 1 and 3 randomly chosen samples (columns 1 and 3). We have not tested more few-shot samples because we exceeded the maximum input length in some cases. Few-shot experiments are run five times, and the average accuracy is reported. ✗ corresponds to experiments that have failed because of limited GPU VRAM (32GB). The results show that the vast majority of the experiments resulted in an accuracy of around 50% which is similar to the majority classifier or random guessing (note that the SWDA rhetorical question test corpus is balanced).

Model	Prompt	0	1	3
Flan-T5-xxl	P1	52.7	-	-
Flan-T5-xxl	P2	**56.3**	52.2	53.4
Flan-T5-xxl	P3	49.2	-	-
OPT-6.7B	P1	52.3	-	-
OPT-6.7B	P2	52.0	51.2	**54.2**
TK-11B	P4	55.9	**57.7**	✗

Table 1. RQ accuracy on the balanced test set of SWDA, for an increasing number of training samples per class: 0 (ZSL), 1 and 3 (FSL).

5 PROPOSED APPROACH

Several parameter-efficient finetuning approaches, such as LOw-Rank-Adaptation and adapters, have been proposed recently to finetune large PLMs in a cost-efficient way. They often obtain comparable performances to full finetuning, even when the main model's parameters are quantized to 4 bits as with qLoRA. We rather investigate the properties of the soft-prompting method because, conversely to the

above-mentioned mainstream parameter-efficient approaches, soft-prompting does not modify the internal parameters of the model, in particular the MLPs in the self-attention layers. The fact that they do not modify the MLP parameters, where information is mainly stored, leaves open the question of the real impact that the soft prompts may have on the embedding space that is build at the output of the transformer: is it possible to build a semantic embedding space with contrastive soft prompting, in a similar way than with S-BERT? Is it possible to build a joint embedding space that projects multiple domains into the same embedding space?

In this work, we investigate both properties of soft-prompt tuning in the context of the challenging task of rhetorical question detection. We thus propose a contrastive approach for soft-prompt tuning with the objective of building a semantic embedding space for rhetorical questions, and then exploit it for few-shot instance-based detection of rhetorical questions. We further investigate whether training this soft prompt contrastively on multiple domains/corpora enable building a joint embedding space across domains/corpora. For reproducibility, we provide a GIT repository[4]. We ran our experiments with relatively small uni- and bidirectional representatives of PLMs (namely smallest *gpt2* with 124M parameters and *bert-base-cased*), similarly as in the experiments of [17]. We expect that, by demonstrating both these properties of soft-prompting on small models, these properties shall remain valid on larger models, while the other way is more problematic because similar properties of PLMs are known to only emerge when the PLM is large enough, such as in-context learning, chain-of-thoughts...

5.1 Prompt-tuning – Model Details

In all experiments, we use the BERT [33] pre-trained model (*bert-base-cased*). It was the most robust model in terms of hyper-parameters settings while obtaining good results with relatively low computational costs, compared to the other models we tested, such as GPT-2, Bloom, and Flan-T5 (see Appendix B).

The model parameters are frozen and we train only a final classification layer and 10 new embedding vectors (the soft prompts P_1 to P_{10}) based on preliminary experiments (see Appendix C). These trainable parameters are shown in green in Figure 2.

To preserve the CLS token pre-trained position, the soft prompts P_1 to P_{10} are systematically appended at the end of every input sequence due to the positional encoding. We use padding tokens and an adequate attention mask in order to guarantee the same position of the soft prompts for each input. The input text $T_1 \dots T_n$ is composed of the current question without context, to facilitate comparison across corpora. The last linear layer has one output neuron for binary classification.

[4] `https://github.com/balounjosef/SoftPromptRQ`

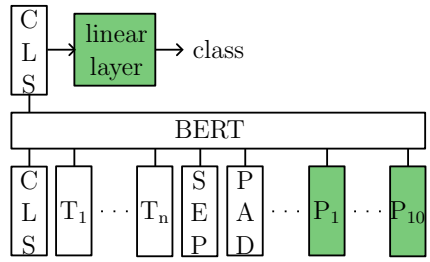

Fig. 2. Soft-prompt model, trainable parameters are filled in green

5.2 Leveraging More Corpora

An important aspect of our approach is that we are using more than one corpus to take advantage of similar domains/tasks and also evaluate the generalization of the model.

5.2.1 Zero-shot Cross-corpus Transfer

The simplest way to test generalization across corpora consists in tuning the soft prompts on the training part of one corpus (called the *source* corpus) and evaluating the performance on the test part of another corpus (the *target* corpus). The experiments are presented in Section 6.1. This follows the intuition of transferring knowledge of similar labels in the same domain (e.g., RQs in dialogues). However, this simple approach does not solve the bias problem mentioned before. This is called Zero-shot Cross-corpus Transfer, because no label of the target corpus is used.

5.2.2 Pseudo-few-shot Cross-corpus Transfer

After training the soft prompts on the source corpus, we may further continue training them on the target few-shot samples: we show in Section 6.2 that good results can be obtained. However, this approach is still limited by the fact that continued prompt-tuning involves setting some hyper-parameters, e.g. the number of epochs, typically using a development corpus of the target task. That is why we call this approach "pseudo-few-shot".

5.3 Contrastive Alignment of Embeddings

The above-mentioned solutions are problematic because the hyper-parameters used to fine-tune the prompts on the target few-shot data strongly depends on the target corpus; assuming the best hyperparamaters are found on the source corpus, or another distant dataset, they are likely to induce a large variance of performances when applied on the target few shots. This is why it is not rare to find published works

that use a development corpus of the target domain to find these hyper-parameters, which is not compatible with the few-shot setting [37].

As a solution, we propose to adopt instead a contrastive learning strategy, where we aim at aligning into the same embedding space both the source and the target corpora. Hence, few-shot classification on the target task may be realized within the framework of "instance-based learning" or "prototype-learning", just by comparing the distances in this joint embedding space between the unknown test sample and the few-shot samples (see Section 6.3 and Figure 4).

Let s denote a sequence of tokens and $E : s \mapsto \mathbb{R}^d$ an embedding space. Prompt-tuning on one (source) corpus creates an embedding space E_{source} that is different from the target embedding space E_{target}. Then, continued prompt-tuning progressively aligns E_{source} towards E_{target}.

In our experiment (see Section 6.3), we try to align E_{MRDA} and E_{SWDA} into a single common embedding space. Contrastive learning is a method of choice to train such joint embeddings [36]. Furthermore, once such a joint embedding space is available, a parameter-free instance-based learning strategy can be used to perform few-shot learning by computing the distance between an unknown test sample and the known few-shots in this embedding space, and deciding on the output class with K-nearest neighbors. Figure 3 shows the idea behind of contrastive alignment of embeddings. Let's assume that a source and target corpora contains positive and negative classes (depicted as circles and crosses). The goal is to bring close together the positive samples in both target and source corpora and enable the use of classification based on measuring distances between vectors. Another option would be to train a discriminator model that decides, based on two input vectors across datasets, whether or not they share a class. Such an option, however, is more complicated and requires further training and more samples.

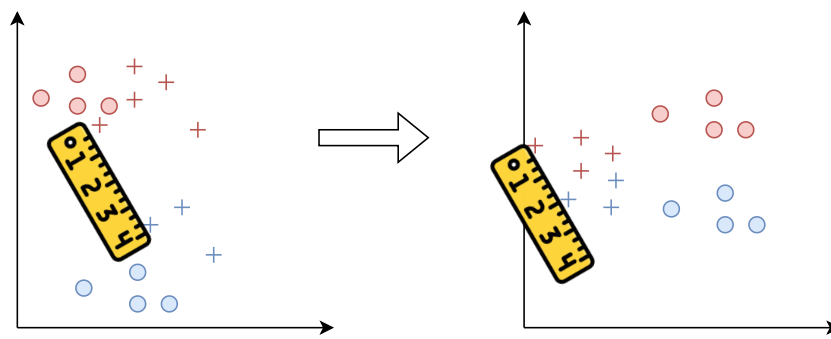

Fig. 3. An illustration of simplified building the join embedding space for two corpora in 2D. Note that Euclidean distance is visualized, but in higher dimensions, the more reasonable option is to choose a cosine distance.

6 EXPERIMENTS

First of all, we provide a comparison of the number of samples used for a particular approach we have tested to solve our research question (see Table 2). An important difference is that the Pseudo-few-shot Cross-corpus Transfer, compared to Contrastive Prompt Tuning, requires additional validation data from the target domain.

Method	MRDA	RQuet	SWDA
Zero-Shot Learning (Sec. 4)	0	0	0
In-context Few-Shot Learning (Sec. 4)	0	0	3
Zero-shot Cross-corpus Transfer (Sec. 6.1)	542	1588	0
Pseudo-few-shot Cross-corpus Transfer (Sec. 6.2)	542	1588	shots + 142 (dev)
Contrastive Prompt Tuning (Sec. 6.3)	542	1588	shots

Table 2. Number of samples used during tuning of each method. The *shots* relates to the number of samples used for few-shot (1, 5, 10 and 25-shots).

6.1 Zero-shot Cross-corpus Transfer

This experiment does not assume any annotated target labels, and directly transfer information from one corpus to another: concretely, prompt-tuning of the pretrained language model is realized independently on each of the four corpora, SWDA, MRDA, RQuet and SarcasmV2, leading to four task-dependent models. Then, every model is directly evaluated on the test part of all corpora. The mapping between the respective four sets of binary labels is shown in Table 3. Cross-corpus accuracies of the BERT model are reported in Table 4.

class	SWDA	MRDA	RQuet	SarcV2
0	non-RQ	non-RQ	ISQ	not sarcastic
1	RQ	RQ	NISQ	sarcastic

Table 3. Mapping between labels in the four corpora

	Test			
Train	**SWDA**	**MRDA**	**RQuet**	**SarcV2**
SWDA	**80.9**	53.9	61.7	48.6
MRDA	*72.7*	**68.6**	55.0	47.5
RQuet	*67.2*	42.2	**75.0**	51.2
SarcV2	42.2	43.1	51.7	**74.4**
SOTA	81.3	–	77.7	–

Table 4. Single-corpus (diagonal) and cross-corpus accuracy [%] of soft-prompt model (*bert-base-cased*). SOTA results: SWDA [10], RQuet [11]. For SarcV2, the authors reports the F1 score 76.0% [27]

Obviously, the best results are obtained when the train and test sets belong to the same corpus (see the diagonal in Table 4).

We can further conclude from these results that:

1. soft-prompt BERT model obtained competitive results with (SOTA);
2. transfer from MRDA and RQuet corpora to SWDA seems possible (see values in italics in Table 4);
3. conversely, transfer from SWDA to MRDA fails as the results are close to random guessing, which may be due to a bias in a corpus;
4. sarcastic and rhetorical questions do not seem to be well aligned.

We focus next on transferring to the SWDA corpus, thereafter called the *target* corpus, while MRDA, RQuet and SarcV2 are the *source* corpora. Since the SWDA is the largest rhetorical question corpus in the collection, the results are more reliable. This choice is also supported by observations and further discussed in Section 7.

6.2 Pseudo-few-shot Cross-corpus Transfer

The cost of producing a few annotated examples for the target task is often moderate, it is reasonable to assume that a small number of labeled samples per target class is available. Exploiting them in our model may be done with *continued prompt-tuning* on the few-shot samples, i.e., simply continuing standard prompt-tuning iterations but on a new set of training samples[5]. Hence, in this experiment, we randomly sample 1, 5, 10 and 25 samples per class from the SWDA corpus. As this sampling may induce variable results, we run this experiment ten times and average the results. We report in Table 5 the results of the following models:

- BERT prompt-tuned from scratch (directly prompt-tuned on the target SWDA few-shots). In this case, there is no transfer and the green trainable parameters in Figure 2 are randomly initialized (the first column with the header None in Table 5);
- BERT prompt-tuned on one of three source corpora (MRDA, RQuet, SarcV2) or their combination (MRDA+RQuet and All) and further continued-prompt-tuned on the target SWDA few-shots. The best results were obtained by first prompt-tuning on the MRDA train corpus and then continued-prompt-tuned on the target SWDA 25 shots.

The learning rate for prompt-tuning is $2e$-4; it has been chosen with preliminary experiments on the MRDA corpus only. However, note that although the model

[5] To the best of our knowledge, we introduce here the term "continued prompt-tuning", which is inspired by the more common term "continued pretraining"

	Pseudo-few-shot Cross-corpus Transfer					
Shots	**None**	**MRDA**	**RQuet**	**SarcV2**	**MRDA+RQuet**	**All**
1	58.8	**73.0**	68.2	43.2	72.7	65.6
5	63.3	**73.9**	71.2	48.6	73.8	67.6
10	64.8	**75.1**	72.5	50.5	73.8	67.3
25	66.6	**76.4**	74.1	51.8	73.9	68.9

Table 5. 10 runs average accuracies [%] on SWDA corpus: we gradually performed continued-prompt-tuning of three models from the previous experiment (Section 6.1 and Table 4)

parameters are only trained on the few-shot samples, we also perform early stopping on the development corpus of SWDA; so the experiment should not be really considered as few shot-learning, which explains the term "pseudo-few-shot" in the section's title. This is a generic issue with few-shot learning experiments, which is usually "solved" with one of the two following approaches: either use a development corpus in addition to the few-shots (as we did), or further split the few-shots into a train and development corpus. As discussed in [37], none of these options is satisfying: the first one because more than a few data points are used to train the model; and the second one because setting hyper-parameters with a standard tuning procedure on so little data inevitably induces a significant variance of the error, and in most concrete use cases, the few-shots are given, and there is no way to tell whether performances will be far from the average or not. We propose a real few-shot learning approach in Section 6.3.

Continued prompt-tuning from both the MRDA and RQuet models on the target few-shots always improves results. However, initial prompt-tuning of BERT on the combined MRDA and RQuet corpora does not help, as compared to simply prompt-tuning BERT on MRDA. Similarly, initial prompt-tuning of BERT on SarcV2 is useless, as the accuracy stays close to a random guess.

This experiment shows that, just like fine-tuning, prompt-tuning may also be successfully used to transfer information from related tasks when applied in the few-shot setting, and that this approach may be useful for the rhetorical question detection task, which is difficult to handle for large pre-trained models as discussed in Sec. 4.

6.3 Contrastive Prompt Tuning – Alignment of Embeddings

In this scenario, we replace the final classification layer in Figure 2 by a projection layer that reduces the 768-dimensional BERT output to a 32-dim embedding. This dimensionality reduction is useful due to the "curse of dimensionality" challenge, which makes similarity search in high dimensions difficult. The final embedding size of 32 is based on a few trials made on the source corpora. The trainable parameters (10 prompts and final projection layer) are now trained with the triplet loss [35]:

$$\mathcal{L}(A, P, N) = \max[d(f_A, f_P) - d(f_A, f_N) + \alpha, 0] \quad (1)$$

where f is an embedding obtained by our model, d is a distance function in the embedding space and $\alpha = 0.5$ is the margin. The goal of the margin is to reduce overfitting, stabilize a training process and, last but not least, helps the model to generalize better. We experimented on MRDA corpus with different distance functions (Euclidean, Manhattan, Cosine). Similar results were obtained, probably because of layer-normalization in BERT, as the norms of the projected embedding vectors are more or less similar. Finally, we chose the cosine distance.

We create triplets that consist of three distinct samples: *anchor* (A), *positive* (P), and *negative* (N). An anchor and positive sample share the same class (label), while the negative sample is representative of a different class. Optimizing the parameters of the network brings A and P closer in the feature space and pushes A and N farther apart [34].

Figure 4 summarizes the proposed approach. As before, BERT is first prompt-tuned on one or several source corpora, but now with the triplet loss. Then, in the second transfer step, BERT is further prompt-tuned with triplets composed of anchors that come from either a source corpus or the target few-shots, while positive and negative samples are sampled from the target few-shots. Hence, many of the second-step triplets mix samples from the source and target corpora, which consequently creates a joint embedding space.

Another advantage of the proposed contrastive approach is that there is no hyper-parameter (learning rate and early stopping) to tune on a target development corpus: contrastive training is run until convergence[6] and then few-shot classification is achieved with the nearest neighbors method. Thus, this approach achieves "real few-shot classification", as described in [37].

The main objective is to build a joint embedding space for rhetorical questions that is common to all related corpora. First, we discard SarcV2, which is not related enough to the task, as discussed in Section 7. We thus consider both MRDA and RQuet as source corpora and SWDA as the target corpus.

Table 6 compares the accuracy obtained with few-shot K-NN classification, before and after the second step of alignment (the last two columns). The remaining columns on the left are results from Table 5. In all configurations, we set the K to the closest odd number $\leq \sqrt{N}$ where N is the number of shots available, since it is standard practice to deal with outliers. So for the 1, 5, 10, and 25 shots, the K is set to 1, 1, 3, and 5 respectively, but we have also observed that $K = 1$ gives similar results. We can draw several conclusions from Table 6:

1. building a better embedding space with interpretable distances comes at the cost of lower accuracy, as compared to the best pseudo-few-shot learning method

[6] We use randomly generated triplets from training data and measure the ratio of correctly predicted triplets so that positive distance is lower than negative distance: $d(f_A, f_P) < d(f_A, f_N)$.

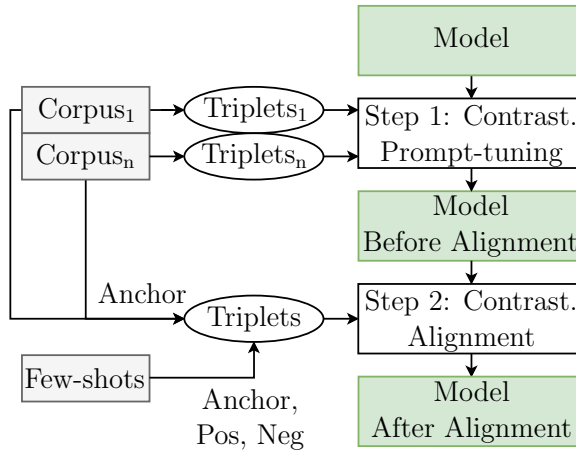

Fig. 4. Outline of the contrastive training strategy

that directly optimizes the accuracy. This is particularly visible in the first-step models. Note however that for these two contrastive models, no SWDA development data is used;

2. aligning the embedding space in step 2 clearly improves classification over the unaligned embedding space in step 1;

3. aligning the embeddings in step 2 requires enough few-shots: the results are poor with only 1-shot, but they tend to get closer to the best pseudo-few-shot model with more shots.

Shots	Pseudo-few-shot MRDA+RQuet	Cross-corpus Transfer MRDA+RQuet	Before alignment MRDA+RQuet	After alignment MRDA+RQuet
	Dev set ✓	Dev set ✗	Dev set ✗	Dev set ✗
1	**72.7**	71.1	60.9	68.7
5	**73.8**	73.3	67.9	73.6
10	73.8	73.6	69.1	**74.8**
25	73.9	73.6	68.7	**75.1**

Table 6. Overall summary; 10 runs average accuracies [%] on SWDA: the first column is the continued-prompt-tuning of pre-trained BERT prompt-tuned on the specified data (Section 6.2); the next column is the contrastive model with independent triplets per corpus; the last column is the contrastive model after the alignment step with mixed triplets across corpora (both covered in Section 6.3). Note that Dev set ✓and Dev set ✗ indicates whether or not the method requires target development set (as already indicated in Table 2).

7 DISCUSSION

Although the best absolute performance is obtained with the pseudo-few-shot model, the comparison with the joint embeddings model is not totally fair, as the latter does not use any development corpus from the target task. Furthermore, building a joint embedding space in a contrastive way is more interesting as the distances between samples are interpretable, and information coming from the three corpora is encoded into this unique common space, which might be in the future enriched with new related tasks and corpora.

Despite our initial intuition, that sarcastic questions may often also consist of rhetorical questions, our attempts at merging both tasks failed experimentally. We analyze next qualitatively the relations between the embeddings space of each corpus independently obtained after the initial prompt-tuning phase.

7.1 How to Interpret Visualizations?

Fig. 5 and 6 show the visualizations of embedding spaces. In this way, we try to plot the relations between corpora and the quality of our "transformation" provided by the contrastive training strategy. Each of the figures contains two or three circles. Each circle represents a particular corpus. The points within represents either a positive or negative class sample. The samples are classified based on cosine distance therefore only the angle matters. The origin is in the center of the circle.

The most important is the outer circle since it is related to our target corpus (SWDA). If all corpora (or their embedding spaces more precisely) are well-aligned it can be seen in the image that all the red marks will occur in one area of the circle (within spherical coordinates) and all the blue ones will be in the opposite area across all the circles.

For the visualization of corpus alignments given a specific model, we take the same number of class 0 (red) and class 1 (blue) samples for each corpus. Since we use cosine distance for the training and evaluation phase, we do not care about the norms of the embedding vectors and normalize them. Then, we reduce the dimensionality to 2 utilizing PCA. Since we are interested in the angles, we further re-normalize the vectors so that each concentric circle represents a particular corpus. The classes are distinguished in the same way and also by a different color. For visualization purposes, we further add a small portion of noise to visualize the density. In this way, the alignment between corpora should be visible since the red and blue colored symbols creates "clusters" across circles.

7.2 Analysis

In Figure 5, the contrast between left and right images is striking: in the left image, the red (non-RQ and ISQ) samples of both RQuet and SWDA are mainly located on the left, while the blue (RQ and NISQ) samples of both corpora are mainly located on the right. So both corpora look pretty well aligned.

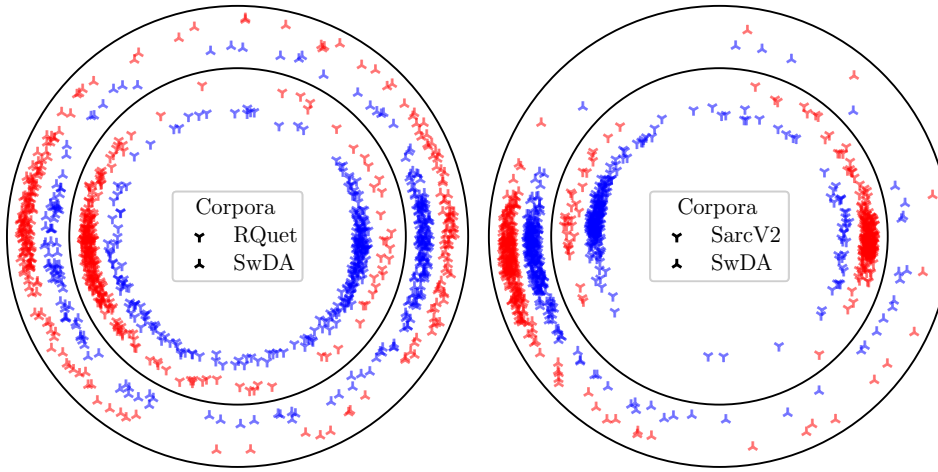

Fig. 5. Vizualization of the model trained on RQuet (left) and SarcV2 (right) before alignment

Conversely, the right part of Figure 5, the red (non-RQ and not sarcastic) samples of SWDA and SarcV2 seem opposed, while the blue (RQ and sarcastic) samples seem more aligned on the left; even worse, the two main PCA-dimensions do not enable to distinguish between RQ and non-RQ from SWDA. So the respective embeddings spaces of SWDA and SarcV2 are not aligned at all, which might explain the poor performances obtained when trying to transfer information from SarcV2

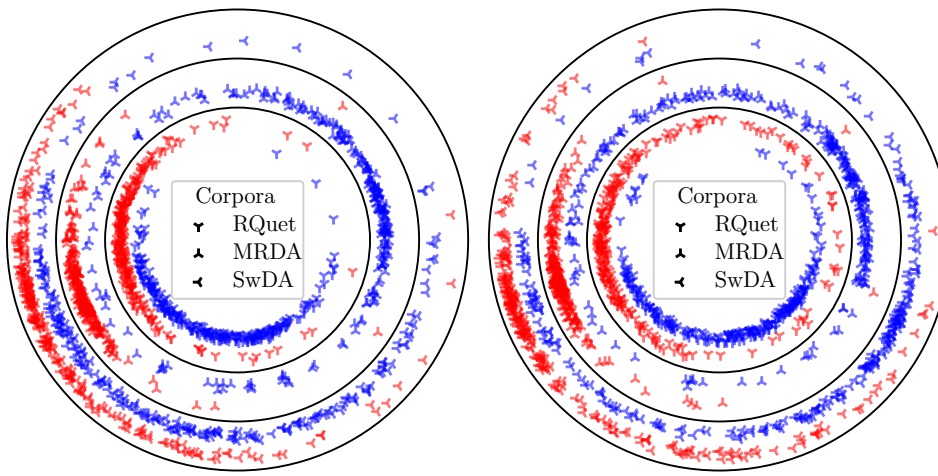

Fig. 6. Vizualization of the model trained on MRDA+RQuet before (left) and after (right) alignment with 25-shots from SWDA

to SWDA.

Note that these visualizations might be used to answer the difficult question of deciding whether a new task is close enough to the target rhetorical question detection task to be considered and merged within the common embedding space built by the three corpora: MRDA, RQuet, and SWDA.

We can further see in Figure 6 the impact of aligning the embeddings space contrastively: if we compare the right image to the left one, the blue dots seem better aligned across the three corpora, which confirms our intuition about the positive effect of aligning the models with cross-corpus triplets.

8 CONCLUSION

In this work, we have shown two new properties of soft-prompt tuning that appear even for relatively small-sized pretrained language models: first, the fact that, just like with full finetuning, using a contrastive training loss builds a semantic embedding space, which is not intuitive given that no internal parameter of the model is modified; mere soft prompts thus seem to be able to more strongly alter the resulting embedding space than expected. The second property is that, also just like with full finetuning, soft-prompts may build a joint embedding space where multiple corpora and labels can be aligned semantically.

We addressed a task, rhetorical question detection, which is difficult to achieve in a zero-shot setting even with large pre-trained language models, as we observed both in the literature on this topic and with our own preliminary experiments (Sec. 4). An additional challenge comes from the fact that when fine-tuning or prompt-tuning such models on one or several of the few existing dialogue corpora annotated with rhetorical questions, the relatively good performances obtained in the training conditions hardly transfer to another corpus with a different application context.

The proposed contrastive soft-prompting strategy (see Figure 3) enables to partly solve both challenges by transferring information from several corpora into the same joint semantic embedding space, using only few shots from the target task. When evaluated in a *real few-shot* setup, our approach obtained better results than 5 out of 6 prompt-tuning models that exploit a much larger development corpus for hyperparameter tuning and proved competitive against the best prompt-tuning model.

Finally, we visualize and analyze the resulting embedding spaces, and suggest a qualitative criterion to select appropriate related corpora that may be further included in this common embedding space.

Acknowledgments

The work of Josef Baloun has been supported by the Grant No. SGS-2025-022 – New Data Processing Methods in Current Areas of Computer Science. The work of Jiří Martínek and Pavel Král has been supported by the project R&D of Technologies for Advanced Digitalization in the Pilsen Metropolitan Area (DigiTech)

No. CZ.02.01.01/00/23_021/0008436. The work of Christophe Cerisara has been supported by the project ANR LLM4ALL. Computational resources were partly provided by GENCI-IDRIS (Grant 2023-AD011011668).

REFERENCES

[1] Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T. & Grefenstette, E.: Large language models are not zero-shot communicators. *ArXiv Preprint ArXiv:2210.14986.* (2022)

[2] Lester, B., Al-Rfou, R. & Constant, N.: The power of scale for parameter-efficient prompt tuning. *ArXiv Preprint ArXiv:2104.08691.* (2021)

[3] He, Y., Zheng, S., Tay, Y., Gupta, J., Du, Y., Aribandi, V., Zhao, Z., Li, Y., Chen, Z., Metzler, D. & Others: Hyperprompt: Prompt-based task-conditioning of transformers. *International Conference On Machine Learning.* pp. 8678-8690 (2022)

[4] Vu, T., Lester, B., Constant, N., Al-Rfou, R. & Cer, D.: Spot: Better frozen model adaptation through soft prompt transfer. *ArXiv Preprint ArXiv:2110.07904.* (2021)

[5] Khan, Z.—Fu, Y.: Contrastive Alignment of Vision to Language Through Parameter-Efficient Transfer Learning . *The Eleventh International Conference On Learning Representations* . (2023), https://openreview.net/forum?id=x0BPR9iXc1

[6] Paul, S., Hong, L. & Chi, E.: What is a question? Crowdsourcing tweet categorization. *CHI 2011.* (2011)

[7] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. & Carvey, H.: The ICSI meeting recorder dialog act (MRDA) corpus. (International Computer Science Inst Berkely CA,2004)

[8] Godfrey, J., Holliman, E. & McDaniel, J.: SWITCHBOARD: Telephone Speech Corpus for Research and Development. *Proceedings Of The 1992 IEEE International Conference On Acoustics, Speech And Signal Processing - Volume 1.* pp. 517-520 (1992)

[9] Zhuang, Y. & Riloff, E. Exploring the Role of Context to Distinguish Rhetorical and Information-Seeking Questions. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics: Student Research Workshop.* pp. 306-312 (2020)

[10] Bhattasali, S., Cytryn, J., Feldman, E. & Park, J.: Automatic identification of rhetorical questions. *Proceedings Of The 53rd Annual Meeting Of The Association For Computational Linguistics And The 7th International Joint Conference On Natural Language Processing (Volume 2: Short Papers).* pp. 743-749 (2015)

[11] Kalouli, A., Kehlbeck, R., Sevastjanova, R., Deussen, O., Keim, D. & Butt, M.: Is that really a question?: Going beyond factoid questions in NLP. *14th International Conference On Computational Semantics: IWCS 2021.* pp. 132-143 (2021)

[12] Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E. & Walker, M.: Creating and characterizing a diverse corpus of sarcasm in dialogue. *ArXiv Preprint ArXiv:1709.05404*. (2017)

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & Others: Language models are few-shot learners. *Advances In Neural Information Processing Systems*. **33** pp. 1877-1901 (2020)

[14] Zhao, Z., Wallace, E., Feng, S., Klein, D. & Singh, S.: Calibrate before use: Improving few-shot performance of language models. *International Conference On Machine Learning*. pp. 12697-12706 (2021)

[15]

[16] Gao, T., Fisch, A. & Chen, D.: Making pre-trained language models better few-shot learners. *ArXiv Preprint ArXiv:2012.15723*. (2020)

[17] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z. & Tang, J.: GPT understands, too. *ArXiv Preprint ArXiv:2103.10385*. (2021)

[18] Li, X. & Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. *ArXiv Preprint ArXiv:2101.00190*. (2021)

[19] Dayal, V.: Questions. (Oxford University Press,2016)

[20] Kimball, J., Adams, D., Campbell, M., Cohen, V., Lovins, J., Maxwell, E., Nygren, C. & Reighard, J.: Papers from the 7th Regional Meeting of the Chicago Linguistic Society. (University of Chicago, Chicago Linguistic Society,1971)

[21] Schmidt-Radefeldt, J.: On so-called 'rhetorical'questions. *Journal Of Pragmatics*. **1**, 375-392 (1977)

[22] Frank, J.: You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal Of Pragmatics*. **14**, 723-738 (1990)

[23] Gutiérrez Rexach, J.: Rhetorical questions, relevance and scales. *Revista Alicantina De Estudios Ingleses, No. 11 (Nov. 1998); Pp. 139-155*. (1998)

[24] Schaffer, D.: Can rhetorical questions function as retorts?: Is the Pope Catholic?. *Journal Of Pragmatics*. **37**, 433-460 (2005)

[25] Martinek, J., Cerisara, C., Kral, P., Lenc, L. & Baloun, J.: Weak supervision for Question Type Detection with large language models. *INTERSPEECH 2022-*. (2022)

[26] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv Preprint ArXiv:1910.13461*. (2019)

[27] Oraby, S., Harrison, V., Misra, A., Riloff, E. & Walker, M.: Are you serious?: Rhetorical questions and sarcasm in social media dialog. *ArXiv Preprint ArXiv:1709.05305*. (2017)

[28] Ranganath, S., Hu, X., Tang, J., Wang, S. & Liu, H.: Identifying rhetorical questions in social media. *Proceedings Of The International AAAI Conference On Web And Social Media*. **10**, 667-670 (2016)

[29] Zhao, Z. & Mei, Q.: Questions about questions: An empirical analysis of information needs on twitter. *Proceedings Of The 22nd International Conference On World*

Wide Web. pp. 1545-1556 (2013)

[30] Chung, H., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S. & Others: Scaling instruction-finetuned language models. *ArXiv Preprint ArXiv:2210.11416*. (2022)

[31] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. & Others: Opt: Open pre-trained transformer language models. *ArXiv Preprint ArXiv:2205.01068*. (2022)

[32] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A., Arunkumar, A., Stap, D. & Others: Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Proceedings Of The 2022 Conference On Empirical Methods In Natural Language Processing*. pp. 5085-5109 (2022)

[33] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)

[34] Balntas, V., Riba, E., Ponsa, D. & Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks.. *Bmvc*. **1**, 3 (2016)

[35] Hoffer, E. & Ailon, N.: Deep metric learning using triplet network. *International Workshop On Similarity-based Pattern Recognition*. pp. 84-92 (2015)

[36] Wang, T. & Isola, P.: Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *Proceedings Of The 37th International Conference On Machine Learning*. (2020)

[37] Schick, T. & Schütze, H.: True Few-Shot Learning with Prompts—A Real-World Perspective. *Transactions Of The Association For Computational Linguistics*. **10** pp. 716-731 (2022)

[38] Suhas Ranganath: Identifying Rhetorical Questions in Social Media. *Proc. Of AAAI ICWSM*. (2016)

Appendix A CORPORA STATISTICS

The number of samples and the input size in terms of input tokens for *bert-base-cased* model are provided in Table A1 and Table A2, respectively.

Corpus	**train**	**dev**	**test**
SWDA	560	142	256
MRDA	476	66	102
RQuet	1270	318	180
SarcV2	5216	652	652

Table A1. Number of samples in train, development and test part of each balanced corpus

Corpus	Max	Mean	Median	90-perc.
SWDA	46	13.36	12	22
MRDA	65	10.69	9	20
RQuet	117	20.27	15	39
SarcV2	411	60.85	48	124

Table A2. The number of input tokens in samples from given corpus with *bert-base-cased* tokenizer (maximal, mean, median and 90 percentile value)

Appendix B OTHER MODEL RESULTS

The results of different prompt-tuning models on the full SWDA corpus are provided in Table B1.

Model	Prompts	Learnable Parameters	SWDA Acc
bert-base-cased	10	8449	80.9
bert-base-cased	50	39169	81.2
bert-large-cased	10	11265	81.2
gpt2	10	8449	78.9
google/flan-t5-large	10	11265	75.4
google/flan-t5-large	50	52225	80.1
bigscience/bloom-7.1b	10	45057	80.5

Table B1. Results of different prompt-tuning models on full SWDA corpus

The zero-shot cross-corpus results of gpt2 soft-prompt model are provided in Table B2. Compared to Table 4 from Sec 6.1 where the *bert-base-cased* was employed, the results are generally worse except SarcV2. We also noticed better results when transferring from SWDA to MRDA and RQuet which indicates that the transferability across corpora may also depend on the model. In that case, the proposed qualitative criterion is beneficial. It may help to decide whether to include the corpus given the concrete model or not.

	Test			
Train	**SWDA**	**MRDA**	**RQuet**	**SarcV2**
SWDA	**78.9**	60.8	64.4	46.6
MRDA	67.6	**66.7**	48.9	41.7
RQuet	67.2	48.0	**71.1**	50.6
SarcV2	45.3	45.1	53.3	**78.1**

Table B2. Single-corpus (diagonal) and cross-corpus accuracy [%] of gpt2 soft-prompt model.

Appendix C NUMBER OF PROMPTS

The preliminary experiments to decide the number of prompts are provided in Table C1 and Table C2.

Prompts	**RQuet**	**MRDA**	**SWDA**	**AVG**
1	72.2	68.6	76.2	72.3
5	75.6	67.6	75.8	73.0
10	75	68.6	80.9	74.8
20	73.9	66.7	80.5	73.7
50	69.4	66.7	81.2	72.4

Table C1. Results of prompt-tuning *bert-base-cased* model with different number of prompts on full corpus

Prompts	**SWDA + left context**
1	77.8
10	81.1
50	73.9

Table C2. Results of prompt-tuning gpt2 model with different number of prompts on full SWDA corpus utilizing the left context