Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3780–3787
Marseille, 20-25 June 2022

Decorate the Examples: A Simple Method of Prompt Design for Biomedical Relation Extraction

Hui-Syuan Yeh, Thomas Lavergne, Pierre Zweigenbaum
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
{yeh, lavergne, pz}@lisn.fr

Abstract

Relation extraction is a core problem for natural language processing in the biomedical domain. Recent research on relation extraction showed that prompt-based learning improves the performance on both fine-tuning on full training set and few-shot training. However, less effort has been made on domain-specific tasks where good prompt design can be even harder. In this paper, we investigate prompting for biomedical relation extraction, with experiments on the ChemProt dataset. We present a simple yet effective method to systematically generate comprehensive prompts that reformulate the relation extraction task as a cloze-test task under a simple prompt formulation. In particular, we experiment with different ranking scores for prompt selection. With BioMed-RoBERTa-base, our results show that prompting-based fine-tuning obtains gains by 14.21 F1 over its regular fine-tuning baseline, and 1.14 F1 over SciFive-Large, the current state-of-the-art on ChemProt. Besides, we find prompt-based learning requires fewer training examples to make reasonable predictions. The results demonstrate the potential of our methods in such a domain-specific relation extraction task.

1. Introduction

With the rapid growth of biomedical textual resources in scientific articles, clinical notes, patient forums, social media, and so on, helping humans quickly grasp the key information out of vast content has become necessary. Natural Language Processing and more specifically Information Extraction (IE) algorithms support readers by transforming unstructured text into structured information of interest. Relation extraction (RE), as one of the most important IE tasks, focuses on recognizing the relation types between two entities mentioned in a given sentence (e.g., given *Alfred Hitchcock directed Psycho*, identify that the relation between (*Alfred Hitchcock*, *Psycho*) is *DirectorOf*).

The current state of the art in information extraction is obtained by Transformer models such as BERT (Devlin et al., 2019). Great success has been obtained by adapting BERT architectures to biomedical tasks by additional training (BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019)), or by pretraining from scratch (SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2020)) on biomedical text corpora. More recently, BioMegatron (Shin et al., 2020a) studied the pretraining settings better for the biomedical BERT models; CharacterBERT (El Boukkouri et al., 2020) enabled word representations without requiring segmentation into a priori word pieces, to better represent domain-specific terms in specialized domains. Another stream works on incorporating external knowledge bases into models (Michalopoulos et al., 2021; Liu et al., 2021a).

Compared to previous work that augments training data with biomedical textual or structured data, we explore an alternative training paradigm, *prompting*, to adapt pre-trained models to biomedical RE tasks more efficiently. The current dominant paradigm consists in pre-training a neural model with a language modeling objective such as masked word prediction (Devlin et al., 2019), then fine-tuning this model by retraining it with a different objective related to the target task (e.g., relation extraction). The main idea of prompting, on the other hand, is to keep the language modeling objective as it is, so that pre-trained models can be put to use more directly and efficiently to address the downstream task.

In general, prompting has been shown to be efficient in recent work for a number of downstream tasks (Brown et al., 2020; Schick et al., 2020). Its benefits for domain-specific relation extraction have however received less attention.

The contributions of our paper are as follows:

- We explore prompting on biomedical relation extraction with the ChemProt dataset.
- We present a systematic approach for prompt design in relation extraction tasks for a specific domain without manual effort, including a variety of ranking scores for prompt selection.
- Our results show that prompting boosts model performance, both when fine-tuning on the full training set and in a few-shot training condition.

2. Background

BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and other pre-trained models revolutionized the IE field with universal model designs that are capable of fitting almost all linguistic tasks with minimum change. These models can adapt from pre-training to fine-tuning on various downstream tasks. Thus, the dominant approach for IE tasks nowadays is to adapt these pre-trained language models via objective engineering. However, we can alleviate the gap between the two phases even further by reformulating the fine-tuning tasks into the form of the pre-training task, i.e. masked word prediction. This training paradigm, known as *prompting*, has been proven to be efficient in adapting to downstream tasks in prior work. We refer interested readers to a recent systematic survey on prompting studies (Liu et al., 2021b) for more detail.

The main idea of prompting is to reformulate the given tasks into *templates* with blank positions (e.g., *Steve Jobs left Apple in 1985. Steve Jobs is the ___ of Apple*) and ask a language model to score how well *label words*, i.e., words associated with relations labels, fill these blanks (e.g., *founder*). The majority of earlier work uses only one word to fill the blank, though it is often difficult to accommodate a more complicated relation with a one-label word. (e.g., Relation: *place_of_birth*, with the example *Juan Laporte (born November* 24, 1959*) is a former boxer who was born in Guayama, Puerto Rico.*) On top of that, say we work with binary relations, e.g.

- Relation: *founder*, with examples like: *Steve Jobs left Apple in 1985.*
- Relation: *nationality*, with examples like: *Aragaki Yui is a Japanese actress.*

Applied with the previous template, *Aragaki Yui is the ___ of actress* would not make sense for the newly introduced relation, *nationality*. In practice, most relation extraction tasks are multi-class classification problems which makes the design of templates and the corresponding label words even harder.

Coming up with good templates and label words is the key to good performance. Recent work on prompting contributes various template schemes, for instance for fact probing (Petroni et al., 2019; Jiang et al., 2020a), text classification (Gao et al., 2021), question answering (Khashabi et al., 2020; Jiang et al., 2021), or commonsense reasoning (Trinh and Le, 2018) in the general domain.

Prompting in Relation Extraction Relation extraction (RE) is a classification problem that involves classifying the relation between two entities within or across sentences. It can be binary, multi-class, or multi-label classification and especially for multi-class classification, it is non-trivial to manually design appropriate prompts to distinguish the classes.

Close to (Schick and Schütze, 2021) experimenting with different prompt templates for binary and ternary classification tasks, Chen et al. (2021a) introduced an interpretable and intuitive template for RE to alleviate the required manual effort in a large search space, specifically leaving the label words for human design. An example is *[E1] Google [/E1] is [MASK] [E2] Alphabet [/E2]*, where *[MASK]* represents a blank. Under this template formulation, Han et al. (2021) added extra blanks to fill before the two entities for incorporating entity type information (e.g., *the [MASK] Google [MASK][MASK][MASK] the [MASK] Alphabet*, is expected to be filled as: the *organization* Google *'s parent was* the *organization* Alphabet) and Chen et al. (2021b) presented a synergistic optimization over entity types and relation labels that results in virtual label words.

On another note, Shin et al. (2020b) performed gradient-guided search to automatically search for a suitable template and label words for each class. The resulting prompts are often uninterpretable and inconsistent across relations, hence hard for the models to work with, although their prompt generation is fully automatic. In our work, we start from the template format of (Han et al., 2021) and extend it to a systematic generation of comprehensive prompts without human effort.

Prompting for Biomedical Information Extraction Sung et al. (2021) released BioLAMA, a benchmark composed of biomedical factual knowledge triples for probing biomedical language models. They showed that biomedical language models yield better predictions compared to general models, but they also found that it is due to the model predictions being biased towards certain prompts. To help applying language models with prompting, Anonymous (2021) proposed a method to paraphrase rare words with the help of an extra source (Wiktionary[1]) for natural language inference (NLI) and Semantic Textual Similarity (STS) tasks in the clinical domain.

Prompt-based few-shot learning and fine-tuning have gained attention in the general domain, but is still under-explored in specialized domains. In this paper, we investigate prompting for relation extraction in the biomedical domain.

3. Method

Relation extraction involves identifying the relation type between two entities. We address intra-sentence relations, which are the most frequent in most datasets. For ease of discussion, we will refer to the two entities as e_1 and e_2, which in our case are a chemical and gene respectively. We begin by explaining how we apply prompting to fine-tune language models (Section 3.1.). Next, we move on to the prompt construction (Section 3.2.), introducing the prompt formulation and describing how the examples fed to the models are decorated. Then, we unfold how we come up with the components required for completing the formulation, collecting the candidates for the components (Section 3.3.) and selecting candidates with proposed ranking scores (Section 3.4.). Figure 1 illustrates our method for prompt construction.

3.1. Prompt-based Fine-tuning

Prompting with pre-trained language models can be used for downstream tasks without any explicit training: this is zero-shot prompting. Zero-shot prompting results will simply express the bias of the language models learned from the pre-training corpus. Note that in that setting, training data is still often used for constructing the prompts. We also choose to fine-tune pre-trained language models with prompts, on the full training set and in the few-shot training condition. This is because in biomedical domain prompts, some words can be relatively rare. We refer to this condition as prompt-based fine-tuning.

Specifically, we pass the token representations from the last hidden layer corresponding to the masked input positions, compute similarity with all label word representations, then softmax the similarity scores, as shown in Figure 2 (right). Compared to the conventional fine-tuning that requires a fully-connected layer to process the classification token (see Figure 2, left), the prompt-based fine-tuning we perform does not introduce extra parameters to learn apart from the parameters of the model itself.

3.2. Prompt Formulation

We illustrate below how we prepare examples for the relation extraction task conventionally (1) and with prompting (2).

(1) (Input) *The specificity of tracer uptake was determined by adding the [E1] imipramine [/E1] inhibitor [E2] NET [/E2].*

[1] https://www.wiktionary.org/

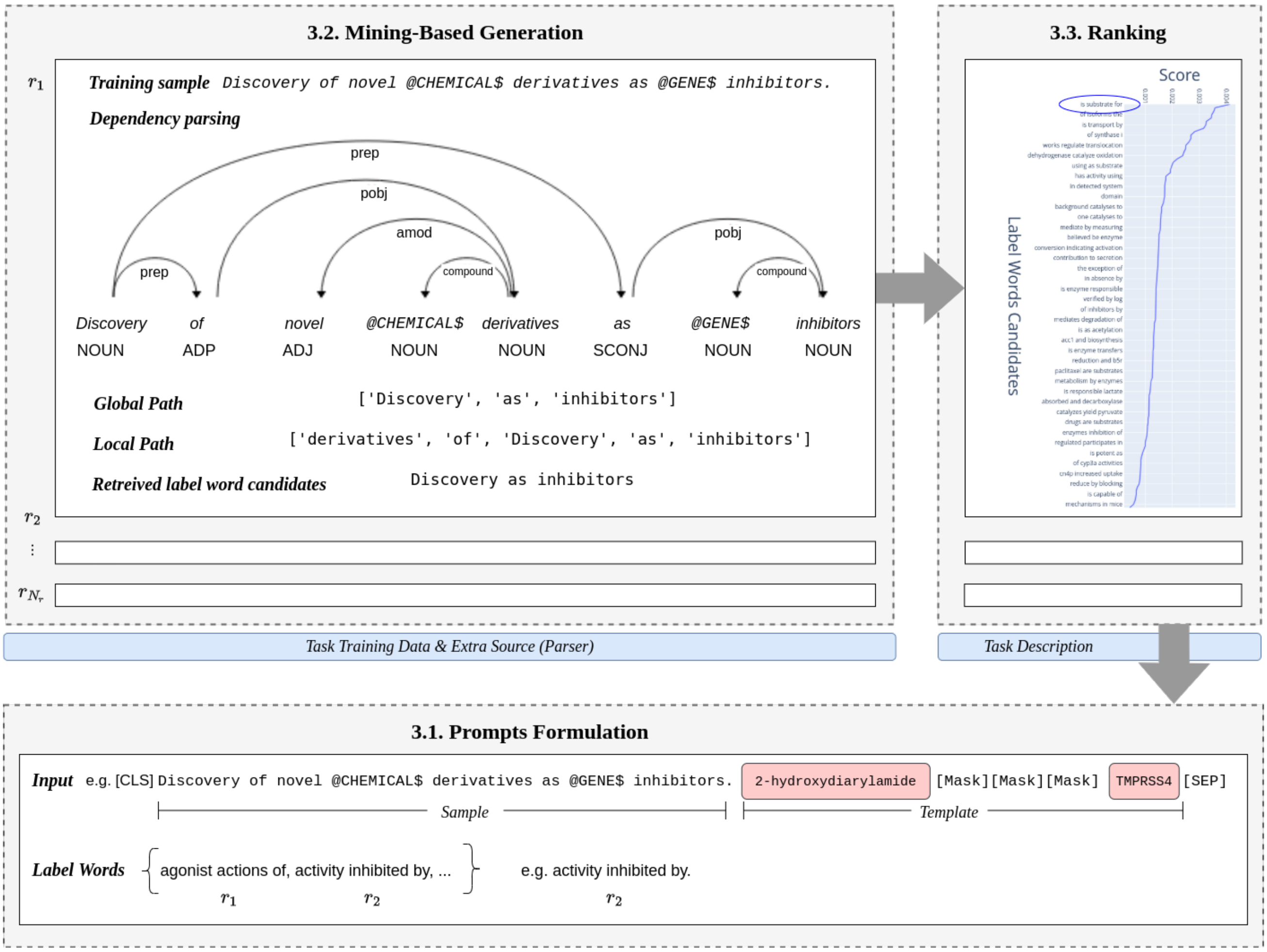

Figure 1: An illustration of the method. Blue marks the resources we use for prompt engineering, red marks the entities.

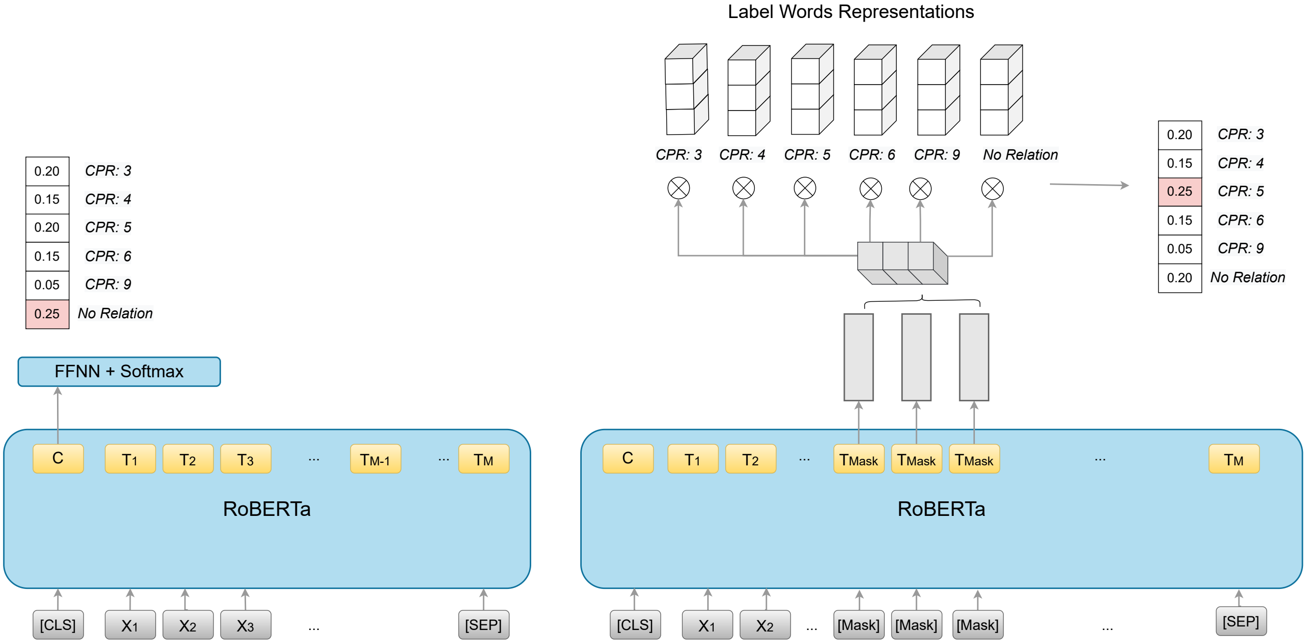

Figure 2: Conventional Fine-tuning (left) vs. Prompt-Based Fine-tuning (right). Compared to the conventional fine-tuning, requiring a fully-connected layer to process the classification token, the prompt-based fine-tuning we take does not introduce extra parameters to learn apart from the parameters of the model itself.

(Label) CPR:4

(2) (Input) *The specificity of tracer uptake was determined by adding the imipramine inhibitor NET. imipramine ______ NET.*

(Label Words) *is inhibitor of*

Following the simple template proposed by (Han et al., 2021), we reformulate each example by appending to it a sentence containing its two entities, with masked tokens between them. In this prompting setting, label words must be defined for each relation. We make room for multiple masked words for better expressiveness, and choose a fixed number of 3 words for simplicity. The model is then expected to score sequences of label words for every relation. The key to model performance lies in choosing relevant label words depending on the task.

3.3. Mining-based Label Word Generation

Toutanova et al. (2015) pointed out that sentences containing synonymous textual relations often share common words, sub-structure, and have similar syntactic dependency arcs. Jiang et al. (2020b) followed that line and used words on the shortest dependency paths between the two entities as label words. This method however often retrieves label words found around the entities rather than between them and hence does not fit our template formulation. Instead, we identify the *local path*: the shortest dependency path from e_1 to e_2 and the *global path* the shortest path from the first word to the last word of a sentence.[2] We take the words appearing on the intersection of *global path* and the *local path* and prune the rest of the words.

3.4. Ranking

To choose the most relevant label words among those mined for each relation r, we score the label word candidates c based upon how salient the word is for the relation. We discuss ranking scores $R(c, r)$ based upon different features. In our notation, $N_c(r)$ is the number of examples labelled r in which candidate c occurs, $N_r(c)$ is the number of relations r in which candidate c occurs, N_R is the total number of relations, $\tilde{c}$ and $\tilde{r}$ are the sentence embeddings for c and for the description of relation r.

Frequency This score directly obtains clues from the training set by checking the number of occurrences.

$$R_{\text{frequency}}(c, r) = N_c(r). \tag{1}$$

Frequency-Specificity The principle is close to tf.idf which suggests that label words that are shared across all relation types are not relevant. This score is defined as:

$$R_{\text{frequency-specificity}}(c, r) = N_c(r) \log \frac{N_R}{N_r(c)}. \tag{2}$$

Similarity This score attempts to take the relation description into consideration (e.g., the relation description for CPR:3 is *activation*, please refer to Table 1 for more details). The frequency score might select irrelevant words that are far from the meaning of the relation type. We use here the cosine similarity between the sentence embeddings [3] of the candidate label words and of the relation description:

$$R_{\text{similarity}}(c, r) = cos(\tilde{c}, \tilde{r}). \tag{3}$$

Combined This score combines the above statistical and semantic properties and is calculated as follows:

$$R_{\text{combined}}(c, r) = R_{\text{frequency-specificity}}(c, r) \cdot R_{\text{similarity}}(c, r). \tag{4}$$

4. Experiments

4.1. Dataset

We use the ChemProt dataset (Kringelum et al., 2016) from the BioCreative VI challenge to investigate the relation extraction task. It contains scientific paper abstracts annotated with 6 relation types between the chemicals and genes in sentences: activation (CPR:3), inhibition (CPR:4), agonist (CPR:5), antagonist (CPR:6), substrate (CPR:9), and no relation. The details are presented in Table 1.
There is no ternary relation or relation associated with more than two entities annotated in the dataset, and the relation is only possible between one chemical and one gene. However, within one single sentence, there can be many annotated relations between different chemical-gene pairs. Also, there exist examples of cross-sentence relations and relations classified with more than one CPR group; but since there are only few of them, we discard these examples and simplify the task into a multi-class problem. With this consideration, we pre-process the dataset into the input format of single sentences, each consisting of the chemical and the gene associated with the assigned label.

Relation	N. train	N. val	N. test
All	19,457	11,820	16,938
Abstracts	1,020	612	800
CPR:3 (activation)	768	550	664
CPR:4 (inhibition)	2,251	1,094	1,661
CPR:5 (agonist)	173	116	195
CPR:6 (antagonist)	235	199	293
CPR:9 (substrate)	727	457	644
No Relation	15,303	9,404	13,483

Table 1: Description of the ChemProt dataset. The table describes the number of examples actually fed into the model, after the pruning described in Section 4.1.

4.2. Model

We conduct experiments with the off-the-shelf Roberta-base[4] and BioMed-RoBERTa-base [5] pre-trained language

[2] We use the spaCy dependency parser, https://spacy.io/api/dependencyparser

[3] We use Sentence-BERT for acquiring the sentence embeddings, https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

[4] https://huggingface.co/roberta-base

[5] https://huggingface.co/allenai/biomed_roberta_base

models. BioMed-RoBERTa-base is continuously pre-trained on scientific biomedical articles based on the RoBERTa-base architecture. Both models have obtained good performance on biomedical domain tasks (Liu et al., 2019; Gururangan et al., 2020) including the relation extraction task we are studying. For the baselines, we add a linear layer on top of the final hidden state of the [CLS] token to pull out the predictions. For the prompting method, we take the outputs of the masked positions from the last hidden layer, then calculate the similarities with the label word embeddings: these similarity scores serve as our model predictions.

4.3. Hyperparameter Settings

We train with 5 epochs with batch size 8. The AdamW optimizer is used with a learning rate of 3e-5, weight decay rate 1e-2, and epsilon 1e-6. For fine-tuning on the whole training split, we report results over 5 random initializations. For few-shot experiments, the performance of learning with few steps can vary significantly depending on the choice of training and validation splits. To mitigate this instability, performance results are averaged over 5 runs on different random seeds to split into training and validation splits. Specifically, k stands for the number of examples we draw from each CPR group; we resample from the pool for the few cases of a relation containing fewer examples than k. Because of the unbalanced distribution of the dataset, some earlier work applies re-sampling, weighting, or simply exclude the dominant class (*no relation*). On the contrary, we do not employ any extra strategy to reshape the distribution, and examine whether the models can cope with it on their own.

4.4. Experiments

Under the RoBERTa architectures, we set up experiments to compare the prompt-based learning (methods that combine prompting with fine-tuning) and the regular supervised learning without prompts, i.e., we add a sequence classification head on top of the pre-trained language models and perform fine-tuning, on both general and biomedical models. In addition, within the prompt-based learning, we set up experiments for different ranking metrics and their counterpart, random pick without any ranking. Lastly, we evaluate on few-shot settings on RoBERTa-base, where we take prompt-based learning with the ranking metric $R_{\text{combined}}(c, r)$, which is the best for fine-tuning on the full training set, and regular supervised learning.

5. Results

5.1. Results on Prompting

Table 2 shows results for fine-tuning on the full training set. Overall, we see that prompting indeed boosts the performance for both models, especially with BioMed-RoBERTa-base achieving the best results 90.09 (sd: 0.08). To the best of our knowledge, the best result so far was 88.95, accomplished by SciFive-Large (Phan et al., 2021), a heavier model based on T5 (Raffel et al., 2020), which we outperform in the present work.

We experiment with label words selected with the proposed ranking scores as well as a random pick from the candidate pool without ranking. The results are displayed in the bottom pane of Table 2. They show that ranking scores does help: especially, RoBERTa-base performs best with prompts generated with $R_{\text{combined}}(c, r)$ and BioMed-RoBERTa-base performs best with $R_{\text{frequency}}(c, r)$. We expected that $R_{\text{combined}}(c, r)$ would be the best ranking scores; however, BioMed-RoBERTa-base might carry some knowledge on the biomedical vocabulary, causing similarity and specificity not to contribute much and frequency to obtain the top results. Note that our F1-score for BioMed-RoBERTa-base without prompting is behind that reported in the source (81.9, sd: 1.0) (Gururangan et al., 2020). This might be due to the different hyperparameter setting and to the relation class weighting. We also look closer into the performance per class for both approaches, focusing on the best performing BioMed-RoBERTa-base model. Tables 3 and 4 show BioMed-RoBERTa-base with conventional and prompt-based fine-tuning respectively. While the

Model	Ranking	Micro F1 (sd)	Macro F1 (sd)
Conventional			
RB	-	80.09 (0.12)	19.23 (0.63)
BioRB	-	76.69 (0.10)	17.20 (0.91)
Prompt-based			
RB	random	88.17 (0.28)	72.08 (0.50)
	frequency	88.12 (0.51)	72.26 (0.60)
	freq-spec	88.35 (0.11)	72.38 (0.68)
	similarity	88.43 (0.38)	73.02 (0.80)
	combined	*88.60 (0.13)*	*74.13 (3.06)*
BioRB	random	89.55 (0.14)	74.79 (0.41)
	frequency	**90.09 (0.08)**	**76.31 (0.23)**
	freq-spec	90.09 (0.15)	76.17 (0.19)
	similarity	89.99 (0.15)	75.64 (0.50)
	combined	89.78 (0.33)	75.65 (0.70)

Table 2: Conventional fine-tuning, and prompt-based fine-tuning with prompts generated by different ranking scores, full training set: micro- and macro-averaged F1-scores (%). For each condition, we report the average and standard deviation over 5 random runs. RB=RoBERTa-base; BioRB=BioMed-RoBERTa-base; freq-spec=frequency-specificity

	P	R	F1	support
CPR:3	0.00	0.00	0.00	664
CPR:4	45.19	93.32	15.47	1661
CPR:5	0.00	0.00	0.00	195
CPR:6	0.00	0.00	0.00	293
CPR:9	0.00	0.00	0.00	644
No Relation	80.23	98.76	88.54	13483
accuracy			79.52	16938
macro avg	20.90	18.02	17.33	16938
weighted avg	68.28	79.52	71.98	16938

Table 3: Conventional fine-tuning, full training set: performance (%) of BioMed-RoBERTa-base, per class and overall. P=precision, R=recall, F1=F1-score

	P	R	F1	support
CPR:3	70.31	67.77	69.02	664
CPR:4	79.75	76.10	77.88	1661
CPR:5	75.43	67.69	71.35	195
CPR:6	84.09	75.77	79.71	293
CPR:9	59.70	61.37	60.52	644
No Relation	93.36	94.27	93.81	13483
accuracy			89.57	16938
macro avg	77.11	73.83	75.38	16938
weighted avg	89.48	89.57	89.51	16938

Table 4: Prompt-based fine-tuning, full training set: performance (%) of BioMed-RoBERTa-base with combined ranking, per class and overall. P=precision, R=recall, F1=F1-score

conventional fine-tuning (Table 3) only predicts the two major relation types CPR:4 (inhibition) and No Relation, the prompting method (Table 4) predicts more diversely and achieves good performance for the minor relation types, e.g. CPR:5 (agonist) and CPR:6 (antagonist).

5.2. Few-Shot Learning on Prompting

In Figure 3, we show our few-shot learning experiments with Roberta-base. We use the ranking metric $R_{\text{combined}}(c, r)$ for prompting. Both approaches start with a high micro-F1 score, but low macro-F1: the predictions are all on the majority class for both approaches. We see that for prompting, a dramatic drop in micro-F1 occurs at $k = 32$, which for conventional fine-tuning occurs later at $k = 128$. This drop is a turning point where the models start to learn meaningful predictions instead of always predicting the major relation type. Having this turning point earlier shows the better behavior of the prompting method. Besides, we observe larger standard deviation for prompting during the performance climbing. This suggests that the prompts work better with certain few-shot example sets than with others. Overall, this result shows that the prompting method obtains faster language-model-based learning of relation prediction, hence makes training more effective on small numbers of examples.

6. Conclusion

In this paper, we investigate prompting for biomedical relation extraction. We propose methods to systematically generate comprehensive prompts to reformulate a relation extraction task. Under a simple prompt template, label word candidates are mined from the training set with the help of a parser, and we propose various ranking metrics to select the best label words representing the relations. Our results show that prompting outperforms the de-facto training paradigm to apply pre-trained models. The results demonstrate the potential of our methods for domain-specific relation extraction tasks. To advance further, there are still many future directions and possible improvements for the approach: (1) as the label words candidate pool can be small, augmenting the pool with knowledge bases and other existing resources, (2) aggregating multiple label words, and (3) mitigating the bias that language models have with label word calibration (Zhao et al., 2021).

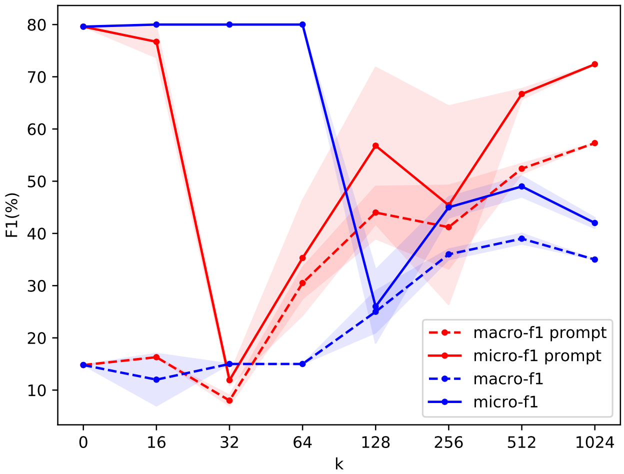

Figure 3: Conventional fine-tuning (in blue), and prompt-based fine-tuning with $R_{\text{combined}}(c, r)$ ranking (in red), few-shot experiments. Micro- and macro-averaged F1-scores (%)

7. Acknowledgements

This research was supported in part by project KEEPHA, ANR-20-IADJ-0005-01, under the trilateral ANR-DFG-JST AI call.

8. Bibliographical References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Anonymous. (2021). Prompt combines paraphrase: Enhancing biomedical “pre-training, prompt and predicting” models by explaining rare biomedical concepts. `https://openreview.net/pdf/24403319293aa6ac3e9304734efee29fb56426eb.pdf`.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chen, X., Xie, X., Zhang, N., Yan, J., Deng, S., Tan, C., Huang, F., Si, L., and Chen, H. (2021a). Adaprompt: Adaptive prompt-based finetuning for relation extraction. *CoRR*, abs/2104.07650.

Chen, X., Zhang, N., Xie, X., Deng, S., Yao, Y., Tan, C., Huang, F., Si, L., and Chen, H. (2021b). Know-Prompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. `https://arxiv.org/pdf/2104.07650.pdf`.

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

El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., and Tsujii, J. (2020). Character-BERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Gao, T., Fisch, A., and Chen, D. (2021). 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)*, pages 3816–3830, Online, August. Association for Computational Linguistics.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., and Liu, X. (2020). Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*. `https://arxiv.org/pdf/2007.15779.pdf`.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Han, X., Zhao, W., Ding, N., Liu, Z., and Sun, M. (2021). PTR: prompt tuning with rules for text classification. *CoRR*, abs/2105.11259. `https://arxiv.org/abs/2105.11259`.

Jiang, Z., Anastasopoulos, A., Araki, J., Ding, H., and Neubig, G. (2020a). X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online, November. Association for Computational Linguistics.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020b). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. (2021). How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., and Hajishirzi, H. (2020). UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November. Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021a). Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, June.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. `https://arxiv.org/pdf/2107.13586.pdf`.

Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., and Wong, A. (2021). UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online, June. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processi ng and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November. Association for Computational Linguistics.

Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., and Altan-Bonnet, G. (2021). Scifive: a text-to-text transformer model for biomedical literature. `https://arxiv.org/pdf/2106.03598.pdf`.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Schick, T., Schmid, H., and Schütze, H. (2020). Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Shin, H.-C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M., and Mani, R. (2020a). BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online, November. Association for Computational Linguistics.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020b). AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Sung, M., Lee, J., Yi, S., Jeon, M., Kim, S., and Kang, J. (2021). Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., and Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal, September. Association for Computational Linguistics.

Trinh, T. H. and Le, Q. V. (2018). A simple method for commonsense reasoning. *CoRR*, abs/1806.02847. `http://arxiv.org/abs/1806.02847`.

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*. `https://arxiv.org/pdf/2102.09690.pdf`.

9. Language Resource References

Kringelum, J., Kjaerulff, S. K., Brunak, S., Lund, O., Oprea, T. I., and Taboureau, O. (2016). Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Appendix

Relations	Random Pick	Frequency	Frequency-Specificity	Similarity	Combined
CPR:3	of src stimulates	is activated by	is activated by	is activated by	is activated by
CPR:4	was difference between	is inhibitor of	design as inhibitors	of inhibition by	activity inhibited by
CPR:5	are gene the	activity is mediated	activity is mediated	agonist actions of	agonist actions of
CPR:6	features of receptor	identified are antagonists	identified are antagonists	known as antagonist	identified are antagonists
CPR:9	was greater in	involved in secretion	involved in secretion	is substrate for	is substrate for
No Relation	effect evaluated in	by concentrations of	by concentrations of	was unable bind	by concentrations of

Table 5: Extracted label words with various ranking metrics