On the Ability to Generalize of Fine-tuned and Few-shot Language Learners for the Relation Extraction Task
COMP 550 Final Report

Kuan Wei and **Zhelin Xu** and **Zedian Xiao**
{kuan.wei, zhelin.xu, zedian.xiao}@mail.mcgill.ca

Abstract

Recently, neural networks based on pre-trained language models are becoming popular for tackling relation extraction task. These models can perform well with a very limited amount of training data. Here in this paper, we compare two different BERT-based models, REDN and BERT-PAIR, that use two different downstream training strategies - fine-tuning and few-shot learning respectively. We aim to understand which strategy can perform better with a limited amount of training data across different relation extraction tasks. We show that the performance of REDN (supervised model) grows much faster than BERT-PAIR (few-shot model) with an increasing amount of training data on SemEval 2010 Task 8, NYT10, and PubMed. Therefore, we believe REDN has a strong potential for solving relation extraction tasks that have a few training data available, while BERT-PAIR can be useful when the amount of training data is extremely scarce.

1 Introduction

Relation extraction, which often appears in the context of information extraction, is a fundamental natural language processing task. Given some input sentence and a pair of entities inside the input sentence, relation extraction seeks to assign the proper semantic relationship between the two words. For instance, if we have the sentence: **Ben visited Hanoi before leaving Vietnam for his next stop in Bangkok**, and we are given the entities **Hanoi** and **Vietnam**. Now the correct relation between Hanoi and Vietnam would be **capital_of**. Now one may ask the question: **What is the capital of Vietnam?** Two plausible entities would be **Hanoi** and **Bangkok**. Here a question answering system would take advantage of the provided semantic relationships to determine that Hanoi is the capital of Vietnam. There has been growing interest for relation extraction in the fields of knowledge base construction, question answering, drug-gene studies and many more.

Before the era of neural models, relation extraction models mostly relied on predefined rules and bootstrapping schemes (Bach and Badaskar, 2007). Through this report, we will focus on the more recent literature which leverages the power of pre-trained language models and few-shot learning. While supervised methods which adapt language models to the relation extraction task achieve state of the art results, they are unable to infer beyond the the provided set of examples and remain quite costly to fine-tune as it requires labelling a large set of examples. This is quite a problem in relation classification as diverse semantic relationships appear in different contexts. As a result, a lot of work has been done in building models that are able to generalize across domains using only a small number of training examples (Baldini Soares et al., 2019), (Gao et al., 2019). In fact, there has also been interest in building models for which no examples are provided (Gong and Eldardiry, 2020), (Levy et al., 2017). One of these models try to capture new relations using additional side information which includes hypernyms and synonyms. Another interesting model seeks to reformulate the relation extraction problem as a question answering problem.

Now we will state the main contributions of this report. We consider both fine-tuning methods which adapt pre-trained language models to relation extraction and few-shot learning models using pre-trained language models. For simplicity, we choose the popular language model BERT and consider a few-shot learning adaptation of BERT, BERT-PAIR, by Gao et al. and a fine-tuned adaptation of BERT which passes outputs from BERT transformer layers to a kernel, by Tian and Li (Gao et al., 2019), (Li and Tian, 2020) . We are interested in the ability to generalize of these two models on unseen data and across domains. We provide a

quantitative analysis on the performance of both models for various sizes of training data for the fine-tuned model and different *K* (number of shots) in the few-shot learning model. We expect the supervised method to perform rather poorly on smaller amounts of data, but to catch up to few-shot models when provided a large quantity of data. Finally, we comment on the effectiveness of few-shot learners as opposed to fine-tuned language models.

In the following sections, we will briefly present BERT, then delve into the compared models and highlight differences in their architecture. Then we will discuss commonly used benchmarks to evaluate these models and describe the formatting of the data sets used for comparison. Finally, we present the experimental procedure and results.

2 Related Work

To our knowledge, no such work has been done before, however some authors adapt their fine-tuned models to evaluate on few-shot benchmarks. In particular Soares et al., in their paper *Matching the Blanks: Distributional Similarity for Relation Learning*, apply their models to both a standard supervised task as well as the few-shot benchmark. They also present results on the performance of their models for increasing task specific tuning data, however they do not make any comparison between fine-tuned and few-shot results (Baldini Soares et al., 2019). We will explore literature dealing with implementing few-shot learning and fine-tuning to natural language processing as some of these provide interesting information about the two methods. Emphasis should be put on the pros and cons of these two approaches when dealing with natural language tasks. Yogatama et al. showed that fine-tuned pre-trained language models still require large amounts of training data to perform well (Yogatama et al., 2019). At that time, there were rather few methods which were based on few-shot learning, however the authors do make the note that they would expect few-shot learners to be able to better generalize on new tasks. Later, Bansal et al. delve into few-shot natural language classification tasks with the intuition that fine-tuning pre-trained models is costly. What they had found is quite interesting, for increasing sizes of pre-trained language models, the performance of few-shot learners based on these pre-trained language models increase. However, they remark that meta-training allows for consistent gains in performance even for small language models (Bansal et al., 2020). Overall, these papers express a multitude of interesting ideas around these methods of training, in this paper, we bring this further and highlight differences between these two methods, when applied to natural language tasks.

3 Literature Review

3.1 BERT

A popular pre-trained language model that has been studied and applied in various areas over the past years is BERT (Devlin et al., 2019). Simply put, BERT follows a transformer architecture and is pre-trained on the unsupervised tasks of predicting a word in a sentence and the likelihood of the next sentence. Due to the vastly available sources of online text data, this large language model does not require any supervision at pre-training. More specifically, it is important to note that BERT is pre-trained on the BooksCorpus as well as the English Wikipedia. The beauty of these pre-trained language models is the fact that we can use them in a transfer learning fashion, that is once they are pre-trained, we can directly load the weights into the BERT architecture and use it on some downstream task, in which case, we say that such model is task-agnostic.

3.2 Learning Paradigms

Before we delve into comparing the fine-tuned model and the few-shot model, it is important to highlight the differences in the learning paradigms. In the standard supervised learning scheme used for the fine-tuned model, we provide a training set of the form: $\mathscr{D}_{train} = \{(x_1, y_1), \ldots, (x_n, y_n)\}, x \in \mathcal{X}$, where $\mathcal{X}$ is the set of all data instances. We use this training set to optimize the parameters of the model θ as to model some function $\hat{y} = f(x; \theta), x \in \mathcal{X}$. In contrast, the meta-learning paradigm that is applied to BERT-PAIR considers a training set of the form: $\mathscr{D}_{meta-train} = \{(\mathcal{D}_1^{train}, \mathcal{D}_1^{test}), \ldots, (\mathcal{D}_n^{train}, \mathcal{D}_n^{test})\}$ where each $\mathcal{D}$ is given by $\mathcal{D}_i^{train} = \{(x_1^i, y_1^i), \ldots, (x_l^i, y_l^i)\}$ and $\mathcal{D}_i^{test} = \{(x_1^i, y_1^i), \ldots, (x_m^i, y_m^i)\}$, which we will refer to as the support set and the query set respectively. An example a support, query set pair can be found in 1. For each sampled instance $(\mathcal{D}_i^{train}, \mathcal{D}_i^{test})$, we pair each data point from the support set to each element to the query set and we determine the probability of whether the two instances belong to the same class.

Support Set	
(a) content_container	Mom put the apple (content) inside the box (container).
(b) component_whole	The handle (component) on the door (whole) is broken
(b) cause_effect	An earthquake (cause) caused the tsunami (effect).
Query Set	
(a) (b) or (c)	You are sick because of COVID.

Table 1: Example of a meta-learning instance

We notice that for the few-shot learning paradigm, each instance may contain examples with different classes from other instances. Hence, such models are said to be *multi-task*.

Various analogies have been made between meta-learning and the way children learn. People argue that standard supervised approaches require a lot of data to perform well, however children seem to be able to easily infer ideas without much supervision. This is exactly the question of interest here. How does the meta-learner compare to a standard supervised model when provided very few examples at training.

For reference, we define a particular few-shot setting by the number of classes in a specific instance and the number of examples per class. If each instance contains 5 examples from 5 different classes, we say that the few-shot model follows a *5-way 5-shot* format.

3.3 BERT-PAIR

The authors of FewRel2.0, one of the benchmarks for few-shot relation classification, proposed BERT-PAIR, a few-shot learning model based on BERT (Gao et al., 2019) [1]. In brief, for every sampled (support, query) set, BERT-PAIR concatenates each sentence from the query set to each sentence from the support set. Now for every concatenated pair, it applies the BERT model for sequence classification to output the probability that the pair shares the same relation. Finally a softmax is applied over the outputs and we infer the relation of the query element as the relation with highest probability.

[1] https://github.com/thunlp/FewRel

3.4 Fine-tuned BERT + Kernel (REDN)

The other model we consider in this analysis is another adaptation of BERT for relation extraction [2]. This model REDN by Li and Tian (Li and Tian, 2020) has seen a considerable amount of success and the authors have claimed that their solution solves various common problems in relation classification such as the existence of a multitude of relations for the same entity pairs. The model consists again of a BERT model and feeds the latter into a kernel. It turns out that they not only use the output from the last layer of BERT but also from the penultimate layer. More specifically, they add the input to BERT E_a with the penultimate layer output of BERT E_w and then apply the kernel over the sum $E_b = E_a + E_w$ and the output from the last layer of BERT E_p. The matrix obtained S_i after applying the kernel is then passed through a sigmoid function after which we obtain the probability that some relation i exists at the position (m, n) in the matrix where m and n are positions of tokens from the input encodings.

4 Datasets

We compare the performance of the two above mentioned models on the following publicly available datasets: NYT, SemEval-2010 Task 8 and PubMed. Furthermore, the FewRel dataset was used for training the few-shot model.

FewRel (2.0). FewRel is a dataset for few-shot relation classification obtained from manual annotation on the Wikpedia corpus and the Wikidata knowledge bases (Han et al., 2018) (Gao et al., 2019).

NYT10. NYT10 contains sentences extracted from The New York Times corpus (Riedel et al., 2010).

SemEval2010 - Task 8. The SemEval-2010 Task 8 dataset was manually extracted from a pattern-based Web search and then labeled by annotators (Hendrickx et al., 2010).

PubMed The PubMed dataset was constructed by the authors of FewRel to test few-shot learners' performance on unfamiliar medical domains.

5 Methodology

In this section, we describe the steps we take in comparing the BERT-PAIR model to the REDN model. The main task is to compare the fine-tuned

[2] https://github.com/slczgwh/REDN

	SemEval	NYT10	PubMed
0.5%	5	N/A	N/A
1%	10	1	1
2%	20	N/A	N/A
3%	30	3	3
5%	50	5	5
10%	100	10	10
15%	150	15	15
20%	200	20	20

Table 2: Number of instances per class in training set

supervised model to the few-shot model for varying training sizes. For each of the above mentioned datasets, we sample a specific amount of examples for 5 specific classes for fair comparison with the few-shot model, these amounts are specified in Table 2. Each of the samples correspond to an approximate percentage of the entire dataset for the 5 selected classes (N/A means that the associated percentage was not taken into consideration). As for the few-shot model, we consider three specific cases: 5-way 1-shot, 5-way 5-shot and 5-way 10-shot. It is important to note that various hyperparameter choices as well as design choices are heavily affected by the available hardware we have access to. All of these models were trained on a google colaboratory instance where 16GB of memory was available on GPUs. Due to this limitation, we are unable to align the number of examples provided for few-shot learning to that of fine-tuning for more than 10 examples per class. For better comparison, future experiments with more than 5 classes and more than 10 examples should be studied.

The BERT-PAIR performances were obtained by training for 30000 epochs on the FewRel data, then evaluated for 1000 iterations on each of the three respective datasets. The batch size used for this was 1 across all three datasets. The REDN model was fine-tuned on the varying sizes of data above. For each of the studied cases, the model was trained until convergence, which occurred after 20 epochs using a batch size of 1 and the model with best performance on the validation set was kept. For each of the cases, we evaluate REDN's performance on the remaining available data. We will compare performances using the accuracy metric.

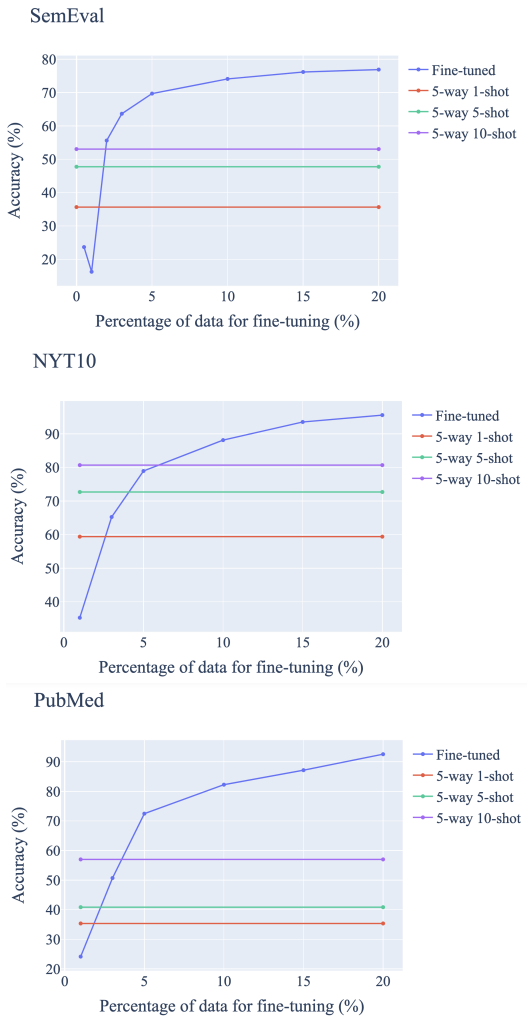

Figure 1: Performances of REDN and BERT-PAIR across datasets

6 Results

In this section, we report the results of BERT-PAIR and REDN on the datasets mentioned above. We randomly sample for each of the described scenarios and average our results.

Figure 1 shows that in general, increasing performance is observed with increasing the amount of training data for both BERT-PAIR and REDN, and BERT-PAIR outperforms REDN only when the amount of training data is very small. In SemEval, even 5-way-1-shot model significantly surpasses the performance of the REDN model with 1% of training data was used. However, when REDN receives more than 2% of the training data, it performs better than the 5-way 10-shot model. In NYT10 and PubMed, when both models have only 1 instance per class for training (1% for REDN versus 1-shot for BERT-PAIR), BERT-PAIR outperforms REDN significantly. When the training

Dataset (N-way, K-shot)	REDN	BERT-PAIR
SemEval (5, 1)	32.25	35.65
SemEval (5, 5)	47.45	47.75
SemEval (5, 10)	55.6	53.05
NYT10 (5, 1)	32.4	59.4
NYT10 (5, 5)	68.6	72.7
NYT10 (5, 10)	92.15	80.7
PubMed (5, 1)	19.6	35.35
PubMed (5, 5)	68.05	40.85
PubMed (5, 10)	85.9	57

Table 3: Few-shot accuracies (%)

data increases to more than 5 instances per class, REDN achieves better results than BERT-PAIR.

A general trend we notice across the three different datasets is the difference in complexity. In fact, both methods seem to perform well on NYT10 and PubMed but SemEval is a more difficult task. We note that this may be influenced by the fact that both REDN and BERT-PAIR utilize BERT for feature extraction. It is possible that BERT provides more accurate understanding on NYT10 and PubMed. However, this hypothesis would need to be further verified.

6.1 What if we modified REDN into a few-shot setting

Due to the overwhelming performance of REDN over a small amount of data, we ask ourselves how REDN would perform in a few-shot setting. What we mean by this is that we randomly sample a support and query set, then fine-tune REDN on the support set and evaluate on the query set. This process is repeated multiple times and we average the performance over the query sets.

It is quite interesting how few-shot performance of REDN is often not only as good but better than BERT-PAIR. This may indicate that REDN is simply a better model than BERT-PAIR overall. For NYT10 and PubMed this method achieves around the same results as the fine-tuned method. However for SemEval this method performs much better. These findings are valuable, as being able to evaluate in the Few-Shot setting not only avoids the use of development set that is used for the fine-tuned model selection, but also reduces the time spent on training.

7 Discussion

By delving deeper into each of the datasets we evaluate on, we notice that the SemEval task was originally formulated to also predict for directionality. That is the order in which the entities are linked together by a semantic relation. This may indicate that to take directionality into account, REDN requires much more examples than BERT-PAIR. Moreover, the fact that REDN is unable to outperform BERT-PAIR on very small number of examples may be due to the fact that REDN contains a much more complex architecture than BERT-PAIR. This is in fact one of the short comings of fine-tuned models, as they often require additional layers to be correctly applied to the desired task.

8 Conclusion

To conclude, we presented two state-of-the-art BERT-based relation classification language models, BERT-PAIR (a few-shot learning model) and REDN (a fine-tuned model), and investigated their ability to generalize on unseen data and across domains. We found that, despite literature gives the credits to few-shot learning models for better generalizations on new unseen tasks, BERT-PAIR only outperforms REDN when the training data contains less than 5 instances per class. This could be due to a superior architecture used by the REDN model. We notice that REDN is able to provide decent classification accuracy with very limited training data. The results demonstrate that fine-tuning pre-trained language model such as BERT with little training data can achieve outstanding performance on various of relation classification tasks.

9 Contributions

Kuan Wei suggested working on relation extraction and mainly contributed to developing and obtaining results. He has also worked on writing of the paper. Zhelin has made an important contributions to the report by explaining datasets and limitations and worked on partitioning datasets and training models. Zedian participated in finding interesting areas to work on and focused on gathering literature for the project, reviewed the code and refactored the code. Finally, his other major contribution was writing various sections of this report.

References

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction.

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

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

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

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. Fewrel 2.0: Towards more challenging few-shot relation classification.

Jiaying Gong and Hoda Eldardiry. 2020. Zero-shot learning for relation extraction.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *CoRR*, abs/1810.10147.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension.

Cheng Li and Ye Tian. 2020. Downstream model design of pre-trained language model for relation extraction task.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, page 148–163, Berlin, Heidelberg. Springer-Verlag.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *CoRR*, abs/1901.11373.