Robust Question Answering with Task Adaptive Pretraining and Data Augmentation

(Stanford CS 224N Default Robust QA Track. Mentor: Rui Wang)

Qitong Cao, Fang Guo
{qitong, fangg}@stanford.edu

ABSTRACT

Existing research suggests that task adaptive pretraining (TAPT) with data augmentation can enhance classification accuracy on a wide array of natural language processing (NLP) tasks. This project aims to evaluate whether TAPT improves performance on a robust question answering (QA) system. The baseline model, which finetunes DistilBERT on SQuAD, NewsQA, and Natural Questions datasets, achieves an EM score of 33.25 and F_1 score of 48.43 when validated on the out-of-sample DuoRC, RACE, and RelationExtraction datasets. Applying TAPT to the out-of-domain unlabeled training datasets using masked language modeling (MLM) without data augmentation, we do not observe an increase in either metric of performance. However, not using TAPT, our model performance is enhanced when we use backtranslations to augment only a small portion of the training data for finetuning, achieving an EM of 36.91 and F_1 score of 50.16 on the out of domain validation set. This model also achieves an EM of 41.63 and F1 of 58.91 on the out of domain test set. These results thus suggest that data augmentation alone, even to a highly limited extent, may account for the improvements in model performance.

Introduction

Most high-performance pretrained language models today are already trained on unlabeled corpora that are massive and heterogeneous. For instance, BERT (Devlin et al. 2018), a commonly used language model, is pretrained on the BooksCorpus and English Wikipedia, which in combination amount to 3.3 billion words encompassing diverse fields of subject matters. The enormity of the language models has led natural language processing (NLP) researchers and practitioners to adopt the routine of finetuning the word representations from such a pretrained language model to the specific task at hand.

However, Gururangan et al. (2020) propose that instead of directly finetuning the language model, *continued adaptive pretraining* will improve performance on a variety of classification-based NLP tasks. The idea is that despite the massive size of a language model, pretraining it on data that are more directly tied to the downstream goal at hand can further enhance performance. In particular, they compare two adaptive pretraining regimes – *domain adaptive pretraining* (DAPT), which pretrains the language model on sufficient data that concerns the specific substantive domains useful for the application at hand, and *task adaptive pretraining* (TAPT), which continues to pretrain a language model on the unlabeled training data tied to the given task – and find that both approaches lead to improved performance on NLP tasks such as relation classification and sentiment analysis.

This paper will focus on evaluating the effect of continued adaptive pretraining on robust question answering (QA), a different NLP task, in an attempt to investigate whether the

improvements observed by Gururangan et al. (2020) can be further extended to non-classification NLP tasks. Specifically, we will adopt the TAPT regime: On the one hand, our "out-of-domain" datasets in our project still use similar passage sources as the "in-domain" datasets (see **Experiments** for details). Consequently, the boundaries between "in-domain" and "out-of-domain" datasets are somewhat vague, making DAPT less applicable. On the other hand, in comparison to DAPT, TAPT is less computationally expensive, using a far smaller pretraining dataset that is more directly relevant to the task at hand. Still, Gururangan et al. (2020) show that TAPT will nevertheless crucial benefit from a larger scale of pretraining corpus through data augmentation: In many of the classification-based NLP tasks evaluated, the differences in performance between the baseline model and one with TAPT are not statistically significant, whereas the improvements from TAPT with data augmentation are significant and substantial. As such, this project will also evaluate whether the improvements can be obtained from data augmentation itself.

Our experiments demonstrate that using TAPT alone without data augmentation does not help improve model performance on robust QA, whereas using data augmentation alone without TAPT leads to a palpable enhancement. Given that our effective results are obtained with a very limited extent of data augmentation, our project suggests that data augmentation may be a crucial factor driving the improvements observed by Gururangan et al. (2020).

Related Work

TAPT is first proposed by Howard and Ruder (2018), who advocate continued pretraining a language model using data directly suited to the downstream task. However, at the time (pre-BERT), language models were generally not trained on massive corpora, and therefore continued adaptive pretraining was more intuitively useful for those pretrained models. Since then, a burgeoning literature (Phang et al. 2018, Sun et al. 2019) have demonstrated the utility of continued supervised training using labeled data that are tied to the end-task on BERT-based language models. Gururangan et al. (2020) propose the routine of TAPT as adopted by our project, i.e., to continue pretraining using masked language modeling (MLM) based on the corpors directly relevant to the end-task, with techniques such as data augmentation to increase the size and diversity of the pretraining data. In terms of applications, in addition to the ones evaluated by Gururangan et al. (2020), TAPT has also been shown to work well for other NLP classification tasks such as offensive language identification (Jayanthi and Gupta 2021) and hostility detection (Raha et al. 2021). By evaluating its performance on robust QA, our project contributes new insights into how effective TAPT can be for NLP tasks not explicitly based on classification.

In addition to TAPT, the other major adaptive pretraining regime is DAPT, first proposed by Alsentzer et al. (2019) and Lee et al. (2019) in the clinical and biomedical domains respectively. Chakrabarty et al. (2020) demonstrate the effectiveness of TAPT for argument mining in online discussions. Since Gururangan et al. (2020)'s comparative study of DAPT and TAPT, there have been NLP applications combining both regimes, such as Zhang et al. (2021)'s hybrid task-oriented dialog system.

Sennrich et al. (2016) propose backtranslation – translating language data to a foreign language using machine translation and then translate the result back to the original language – as a means of data augmentation. The original motivation of this approach is to facilitate improvements in neural machine translation models. The approach can enhance the diversity of the training data, which helps improve the encoding process, as shown by Imamura et al. (2018) In the area of QA, backtranslation has been used to produce high quality training data (Lewis et al. 2019). Our work contributes to this literature by evaluating how backtranslation helps improve QA model performance.

Approach

This project evaluates whether TAPT and data augmentation using backtranslation can improve performance on robust QA. In a QA task, the model takes in a question plus a context paragraph and outputs an answer to the question that is a substring of the given context. (Sometimes, the answer cannot be found in the context, in which case the correct output will be an empty string or a special string indicating N/A.) Robust QA are evaluated on different QA datasets than the ones on which the model is trained in order to demonstrate how generalizable the model can be on out-of-domain queries. See **Experiments** for a detailed description of our in-domain and out-of-domain datasets.

For our baselines, we use the provided *DistilBERTForQuestionAnswering(bert-based-uncased)* model from *Hugging Face*. The first baseline model (`baseline-01`) is selected according to performance on the in-domain validation set, and its evaluative metrics are based on its performance on the out-of-domain validation set. To increase model robustness, we make use of the out-of-domain dataset to evaluate a second baseline model (`baseline-02`), which is selected directly by performance on this out-of-domain validation set. Note that unlike `baseline-01`, the evaluative metrics reported for `baseline-02` will be biased upwards (as an unbiased estimate of the metrics can only be obtained by evaluating `baseline-02` on a held-out test set randomly sampled from the same distribution of the training and validation sets). However, `baseline-02` is still useful because it serves as a foundation for models trained on augmented data.

To facilitate TAPT, we first load the *DistilBERTForMaskedLM(bert-based-uncased)* model and pretrain it on the unlabeled out-of-domain training data. Specifically, we randomly mask 15% of the word tokens in each sentence. If a word is chosen to be masked, we replace it with (1) `[MASK]` token 80% of the time, (2) a random token 10% of the time, and (3) the unchanged original token 10% of the time. This masking procedure is an effective way to mitigate the mismatch between pre-training and fine-tuning (Devlin et al. 2018). After this continued pretraining, we finetune the produced model (`Continue_pretrain`) in the same way the baseline model is finetuned, yielding `TAPT_model`.

For data augmentation using backtranslations, each selected instance (i.e., the question-context-answer triplet) from the training set is first translated into a language randomly selected from a set of 10 commonly used languages, and then the result is translated back into English, using the

BackTranslation 0.3.0 python package that utilizes Google Translate.[1] Note that due to constraints in time and computing resources, we only select the first 1,500 instances of the in-domain training dataset (as well as the entire 381 instances in the out-of-sample training dataset) for backtranslation, as the Google Translate API caps the number of requests it can take in an hour. `BT_model-01` finetunes `baseline-02` with the out-of-domain training data plus the 1,500 backtranslations from the in-domain training data. `BT_model-02` finetunes `baseline-02` with the out-of-domain training data plus the 381 backtranslations from the out-of-domain training data. `BT_model-03` finetunes `baseline-02` with the out-of-domain training data plus both backtranslations. Finally, as a comparison, `BT_model-00` simply finetunes `baseline-02` with the out-of-domain training data without any backtranslations.

Experiments

Data. We use SQuAD (based on Wikipedia), NewsQA (news articles), and Natural Questions (Wikipedia) as in-domain QA datasets, each with 50,000 {question, context, answer} training triplets. We use DuoRC (movie reviews), RACE (exams), and RelationExtraction (Wikipedia) as out-of-domain datasets, each with 127 {question, context, answer} training triplets. These datasets (summarized in the table from project guideline p.6) are used for pretraining DistilBERT (for TAPT), finetuning (for both baseline and TAPT), and evaluation on the validation sets.

Evaluation Metrics. As mentioned in the project guideline, our performance metrics are EM and F_1 scores. EM is a conservative measure where for each question, EM = 1 if the model predicted answer exactly matches the given answer. Because some questions may have correct answers that may not match the provided answer exactly, we also adopt $F_1 = 2(precision \cdot recall)/(precision + recall)$, where precision refers to what proportion of the predicted answer is in the ground truth answer and recall refers to what proportion of the ground truth answer is in the predicted answer.

Experimental Details. As mentioned in the project guideline, the training questions and context paragraphs have been chunked to fit BERT's maximum content size of 512. The loss function is negative log-likelihood (cross-entropy) loss, optimized through the AdamW optimizer. The models are trained with GELU activation function $x\Phi(x)$, where $\Phi(x)$ is the standard Normal cumulative distribution function. Self-attention is implemented with 12 heads, dropout rate 0.1, and 6 layers. Initializer range is set at 0.02. For TAPT masked language modeling, masking probability is set at 0.15.

Experimental Results. Table 1 below summarizes the evaluative metrics for each model. The first baseline model (`Baseline-01`) achieves an EM score of 33.25 and F_1 score of 48.43 on the validation set. Our implementation of TAPT without data augmentation (`TAPT_model`) achieves an EM score of 32.20 and F_1 score of 47.48 on the validation set. In other words, we do not observe an increase in either metric of performance. However, using both the out-of-domain training data and the 1,500 backtranslations from the in-domain training data to finetune our

[1] The ten languages are: Chinese-Simplified, Chinese-Traditional, Japanese, Korean, French, Spanish, Portuguese, German, Russian, and Arabic. Details on the *BackTranslation* package can be found here: https://pypi.org/project/BackTranslation/.

second baseline model, `BT_model-01` achieves an EM score of 36.91 and F_1 score of 50.16 on the validation set, a palpable improvement from the baseline. Additionally, `BT_model-01` model also achieves an EM of 41.63 and F1 of 58.91 on the out of domain test set.

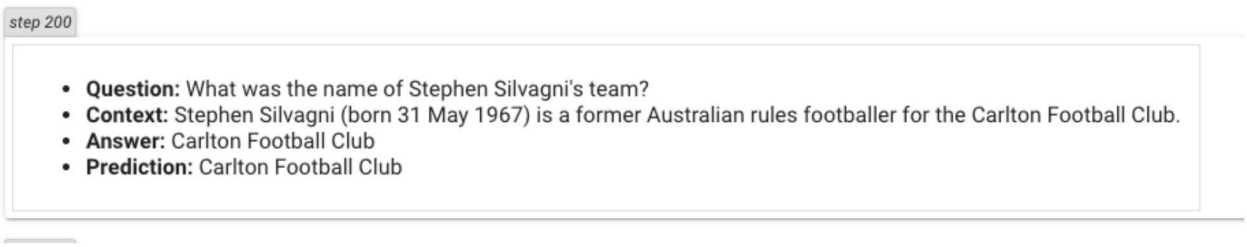
Table 1: Experimental Results

	Out of Domain Validation Set	
	F_1	EM
`Baseline-01`	48.43	33.25
`Baseline-02`	48.59	33.77
`TAPT_model`	47.48	32.20
`BT_model-00`	48.61	32.72
`BT_model-01`	**50.16**	**36.91**
`BT_model-02`	48.65	32.72
`BT_model-03`	50.02	36.13

Analysis

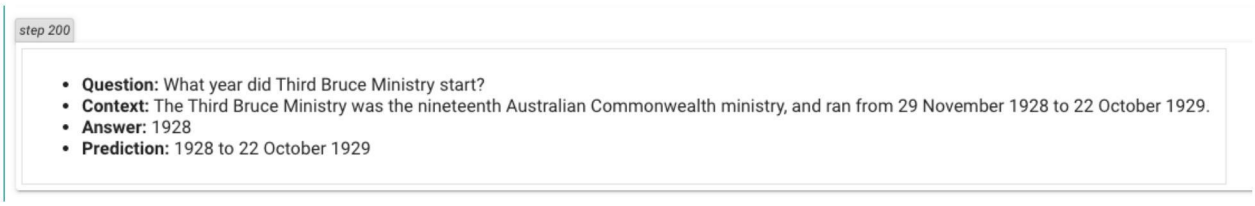
Our results suggest that task adaptive fine training may not be effective for QA when only performed on a small set of additional data. It is possible that applying TAPT to a small dataset that lacks enough diversity may lead to overfitting in the pre-training process, a problem that could be mitigated through data augmentation using backtranslations. This is in line with the results reported by Gururangan et al. (2020). However, even without TAPT, backtranslations alone – even to a very limited extent – palpably enhances our model performance. This outcome thus suggests the great promise of data augmentation through backtranslation, a technique that is relatively easy to implement.

Specifically, we analyzed a few question answering predictions made by our `BT_model-01` on the out of domain validation set:
Example 1:



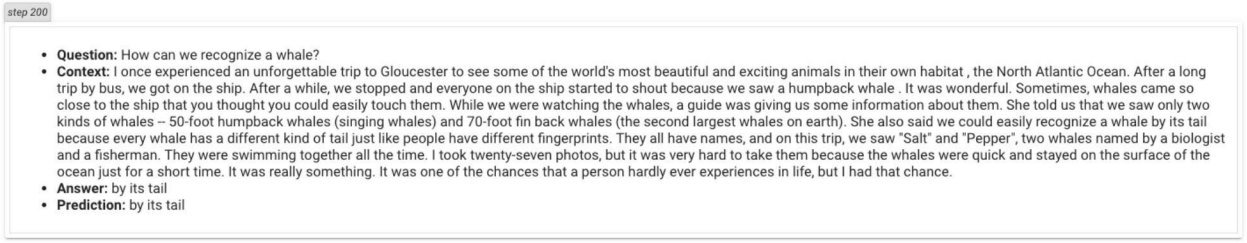
Here, the model makes an accurate prediction, which suggests that the model is able to learn the meaning of "team" and the relation between the person entity "Stephan Silvagni" and "Carlton Football Club".

Example 2:



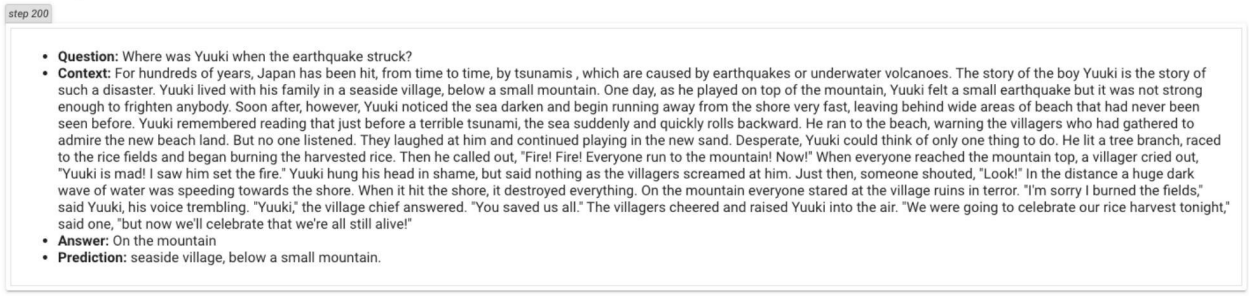
Here, the model's prediction does include the start year "1928", but mistakenly covers the entire time range of the ministry. This indicates that the model is able to recognize year but is unable to understand the specific start and end year.

Example 3:



Despite reading through a long context, the model is able to locate the right answer. However, the sentence with the correct answer contains the same phrase as the questions ("recognize a whale"), which could be the key that helps the model finds the correct answer.

Example 4:



Now, the model makes a wrong prediction. Apparently, the model understands that the question is looking for a location ("where"), but it is unable to find the right location among the many locations that appeared in the context. The model fails because there is no sentence in the context that contains the key phrase in the question ("when the earthquake struck"), so the model cannot perform a simple search like in example 3. Although a human being easily understands that the earthquake struck when Yuki was on the mountain because "he *played* on top of the mountain" and "*felt* a small earthquake", the model struggles at relating "played" and "felt" to the location and the event of earthquake.

It should be noted that due to constraints in time, training data, and computing resources, our project has not addressed the following issues that are important for a rigorous investigation of how much data augmentation can improve model performance on QA and other NLP tasks. First, we are unable to assess whether the improvement in EM and F_1 scores are statistically

significant. Given that the improvements we observe appear substantial, it is likely that the results are not driven by statistical error. Ideally and in future research, however, we will finetune the same model with and without backtranslations using sufficiently many different random samples from the same training datasets (and apply the same backtranslation method to augment the data). Then we will conduct appropriate statistical tests (e.g., Welch's t-test) to compare the means of the performance metrics obtained with and without data augmentation.

Second, given that we have only augmented the first 1,500 instances of the in-domain datasets, we are unsure how much additional improvement data augmentation will result in if pre-trained on more data. More backtranslation-generated data may or may not lead to an approximately linear rate of improvement: It is possible that feeding a few more new sentences will lead to a substantial improvement, in which case we need to focus on more effective methods of data augmentation in order to produce more data for the model to pretrain adaptively. It can also be the case that more data will lead to a diminishing margin of return in performance improvement, in which case feeding a small number of backtranslated sentences may suffice. In future research, we will generate more backtranslations from the in-domain datasets and evaluate multiple models with increasingly more backtranslated sentences for pretraining. Comparing the performance metrics obtained, we will be able to identify an efficient size of data augmentation for finetuning.

Similarly, our results do not necessarily indicate that TAPT does not improve model performance. The same statistical significance issue applies, although again the lack of a difference observed here is not likely to be purely driven by statistical error. Also, it is possible that TAPT and data augmentation combined may lead to more substantial improvements. Future research should look at the potential synergy in this direction. If it happens that TAPT indeed does not help in this particular case, it could be due to the specific language model (DistilBERT) we are using that does not benefit substantially from TAPT, although this is *a priori* quite unlikely, as a smaller language model should theoretically benefit more from continued adaptive pretraining. Finally, it is also possible that robust QA is qualitatively different from classification-based NLP tasks, and adaptive pretraining does not help as much in this (and potentially other tasks that cannot be reduced to classification, such as natural language generation).

Conclusion

This project evaluates whether TAPT or data augmentation using backtranslations can improve model performance for the task of robust QA. The *DistilBERT*-finetuned baseline achieves an EM score of 33.25 and F_1 score of 48.43 on the out-of-domain validation set. Applying TAPT without data augmentation, we do not observe an increase in either metric of performance. However, not using TAPT, our model performance is enhanced when we use backtranslations to augment only 1500 instances of the training data for finetuning, achieving an EM of 36.91 and F_1 score of 50.16. This model also achieves an EM of 41.63 and F1 of 58.91 on the out of domain test set. These results thus suggest that data augmentation alone, even to a highly limited extent, may account for the improvements in model performance. Future research should investigate whether TAPT itself can lead to significant additional effects on performance improvement. For the sake of external validity, it will also be helpful to evaluate whether and

how much TAPT can help other NLP tasks, such as text generation, and perhaps whether similar adaptive pretraining methods can be applied to other areas of artificial intelligence research where transfer learning can be leveraged.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., & Hwang, A. (2020). Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964*.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Imamura, K., Fujita, A., & Sumita, E. (2018, July). Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (pp. 55-63).

Jayanthi, S. M., & Gupta, A. (2021). SJ_AJ@ DravidianLangTech-EACL2021: Task-Adaptive Pre-Training of Multilingual BERT models for Offensive Language Identification. *arXiv preprint arXiv:2102.01051*.

Raha, T., Roy, S. G., Narayan, U., Abid, Z., & Varma, V. (2021). Task Adaptive Pretraining of Transformers for Hostility Detection. *arXiv preprint arXiv:2101.03382*.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Zhang, B., Lyu, Y., Ding, N., Shen, T., Jia, Z., Han, K., & Knight, K. (2021). A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining. *arXiv preprint arXiv:2102.04506*.