Analysis of Indirect Question Answering Models Using Transfer Learning Techniques

Anu-Ujin Gerelt-Od, Lakshmi Menon, Angela Marie Teng

Natural Language Understanding, New York University
{ago265, lsm454, at2507}@nyu.edu
Github: https://github.com/anu-ujin-g/indirectqa_nlu

Abstract

The IndirectQA task aims to understand responses to naturally occurring boolean questions which do not contain direct cue words. Building models that perform well on this task can be instrumental in improving performance of conversational chatbots, as well as interactions with robots or other AI agents. In this paper, we explore the recently developed Circa dataset of indirect question-answer pairs, attempting to replicate and then improve upon its classification results. We first implement the same BERT-based models fine-tuned on other datasets reported in the original paper, and then run similar experiments on other model architectures using T5, RoBERTa, and UnifiedQA. The RoBERTa model fine-tuned on MNLI and Circa achieved the highest accuracy on the test set, in both the strict (87.5%) and relaxed settings (89.6%), as well as the highest F-1 scores on both the strict (86.4%) and relaxed (89.3%) settings.

1 Introduction

Indirect Question-Answering is a natural language understanding task that aims to explore the implicit meaning behind answers to boolean questions. In real world scenarios, not all questions have answers that are easy to interpret, let alone contain direct cue words. Some answers require an understanding of linguistic features or general knowledge in order to be understood, as seen in Example 1. However, not all questions expect a boolean response so some answers, as in Example 2, can be harder to interpret, requiring more than a simple binary classification. This becomes a problem when designing a language model that can be used in applications such as chatbots and voice assistants, with the need to anticipate all types of answers. Researchers at Google published the Circa dataset in October 2020, explained in more detail in Section 3, that explores the IndirectQA task by utilizing pre-trained models and other QA tasks through transfer learning.

The use of various transfer learning techniques consists of making use of a model that was trained on a large-scale dataset, such as BERT, to then fine-tune on a smaller, labeled text dataset, resulting in a much better performance than training on only the latter (Zhang et al., 2020). However, with the abundance of such comprehensive models, it is also costly and time-consuming to train and test all of them for a language task (Dodge et al., 2020). Thus, within the scope of this project, we aim to aid the research in the space of IndirectQA tasks by training on state-of-the-art (SoTA) models to potentially achieve better results. We do this by first replicating the methods from the original paper (Louis et al., 2020) to check for inconsistencies in Section 4.1. Then, we conduct experiments using various pre-trained language models and compare their performance in Section 4.2. The results are shown in Section 5 where we report that the RoBERTa model that was fine-tuned on MNLI and the Circa dataset performed the best across the board.

Q: Want to get some dinner together?
A: I'd rather just go to bed.
Label: No

Example 1: Binary Answer Example from Circa

Q: How was your day?
A: Just a typical Friday.
Label: In the middle, neither yes nor no

Example 2: Non-Binary Answer Example from Circa

2 Relevant Work

Question Answering can be considered to be a form of Information Retrieval (Cao et al., 2010) in which a key task is recognition of answer patterns (Yao, 2014). One type of question is a boolean question, for which the answer is 'Yes' or 'No'. However, early research has found that very often, the answer to such questions does not explicitly contain the words 'yes' or 'no', and is often accompanied by additional speech (Rossen-Knill et al., 1997). With this aim of understanding responses without direct cue words, the recent Circa dataset was created to focus on boolean questions having indirect answers (Louis et al., 2020). In this task, the answers to a question need interpretation and cannot be derived from context alone. Earlier work has attempted to solve this IndirectQA task using Markov Logic Networks (de Marneffe et al., 2009). Green and Carberry (1994), in accordance with Levinson (1983), postulated that a speaker's response provides their evaluation of the propositioned question. This allowed them to use coherence rules as constraints on indirect answers, when building a model to interpret indirect answers. They found the relation between a direct question and an indirect answer to be similar to the relations of "Condition, Elaboration, and Volitional Cause", used in Rhetorical Structure Theory (Mann and Thompson, 1987). In a similar vein, this IndirectQA task is similar to the Natural Language Inference (NLI) task in which a hypothesis is classified as an entailment, contradiction, or neutral response given a premise (Williams et al., 2018), as both the question and answer first need to be interpreted, and then the relation between them must be determined. With this in mind, we believe that we can achieve good performance on this task using models which perform well on NLI tasks, a hypothesis reflected by Louis et al. (2020) as well, in their use of BERT-based models finetuned on BoolQ (Clark et al., 2019) and MNLI (Williams et al., 2018) datasets as a baseline perfomance for the dataset. Models such as T5 (Raffel et al., 2020), RoBERTa (Ott et al., 2019), and UnifiedQA (Khashabi et al., 2020) in particular are current frontrunners on these NLI tasks[1,2]. Improved performance on this topic can be advantageous for use in conversational chatbots or other AI agents, in order to both improve their understanding of human language, and also to produce human-like natural responses. To our knowledge, there are currently no other published works that implement this IndirectQA task using the Circa dataset.

3 Data

The Circa corpus was created from a four-step crowd-sourcing task, with the goal of classifying natural responses as indirect answers. The corpus consists of 3,431 unique questions with up to 10 indirect answers each, for a total of 34,268 question-answer pairs, and is publicly available through Google Research's Github repository[3]. Each question-answer pair has two gold standard labels, one for the 'strict' scheme and one for the 'relaxed' scheme. These labels indicate whether the answer implies *Yes*, *No*, or an in-between classification such as *Probably no* or *Yes, with some conditions*. Upon disregarding labels such as *Other* and *N/A*, the strict scheme has a total of 6 possible classes, while the relaxed scheme has 4 classes (Louis et al., 2020). Sample question-answer-label instances from the dataset are shown in Examples 1 and 2, and the distribution of labels can be seen in Table 3 in the Appendix.

4 Methodology

In the original paper, the IndirectQA task was tested using a BERT model as a baseline, finetuned on the Circa dataset, as well as on a combination of BoolQ+Circa and MNLI+Circa. The best performance resulted from the model finetuned first on the MNLI corpus, and then on the Circa dataset. Performance was measured in terms of accuracy, and this model achieved 84.8% on the strict setting and 88.2% on the relaxed.

To validate the results of the paper and to test the performance of the Circa dataset, we used the same model and analysis methods for replication. We then try to improve these results by utilizing other models in place of standard BERT. Models will be evaluated using overall accuracy and F-1 score, as well as class-wise F-1 scores on a test split of the Circa dataset, following the same setup as the original paper.

[1] https://super.gluebenchmark.com/leaderboard

[2] https://gluebenchmark.com/leaderboard

[3] https://github.com/google-research-datasets/circa

	T5	**RoBERTa**	**UnifiedQA**
Relaxed			
Overall Accuracy	74.7	89.6	89.6
F-1 Score	76.8	89.3	89.2
Yes	79.7	90.9	91.2
Yes, subject to some conditions	87.3	91.0	89.2
In the middle, neither yes nor no	27.2	26.3	41.4
No	77.7	92.7	90.1
Strict			
Overall Accuracy	79.7	87.5	74.7
F-1 Score	84.6	86.4	71.7
Yes	86.6	93.0	80.4
Yes, subject to some conditions	87.9	91.6	86.3
Probably yes / sometimes yes	43.1	36.9	26.9
In the middle, neither yes nor no	27.7	45.3	3.1
Probably no	23.3	22.8	4.8
No	80.3	91.4	73.3

Table 1: Results from experiments using different baseline models

4.1 Replication of Original Results

We replicated the experiments from the matched settings from the original paper, in which the response scenarios are assumed to be seen and the dataset is randomly divided into 60% training, 20% development and 20% testing sets using the following model set-ups: **BERT-YN** (BERT model fine-tuned on Circa), **BERT-BOOLQ-YN** (BERT fine-tuned on BoolQ, then on Circa), and **BERT-MNLI-YN** (BERT fine-tuned on MNLI, then on Circa). All three experiments closely followed the same pre-processing steps and hyperparameters specified in the original paper. The results on the test split of the data are presented in Table 2 in the Appendix.

4.2 Experiments

4.2.1 RoBERTa

Replicating these aforementioned BERT-based models, we achieved similar performance to those obtained by the researchers. Naturally, our hypothesis was that an extension of BERT such as RoBERTa (A Robustly optimized BERT pretraining approach) would perform well on the Circa dataset. RoBERTa (Ott et al., 2019) optimizes for BERT's hyperparameters and training size, and it generally achieves SoTA performance on GLUE, RACE, and SQUAD. Specifically, RoBERTa[4] used a novel dataset, CCNEWS for training, and demonstrated that the use of more data during the pretraining step improves performance on downstream tasks. In addition to pretraining the BERT model on a new dataset, RoBERTa also trains the model longer, with bigger batches, and with more data. Additionally, the RoBERTa model achieves SoTA results on 4/9 of the GLUE tasks, including MNLI. This result encouraged us to test this model and finetune it first on the MNLI dataset, then on Circa.

In both the strict and relaxed settings, the RoBERTa model outperformed the baseline BERT model, achieving an 87% test accuracy on the strict setting and a 91% test accuracy on the relaxed setting. From the class-wise F-1 scores, we see that RoBERTa pretrained on MNLI and Circa performs well on the relaxed classes, with particularly high F-1 scores on the *No* class, achieving 91.0%. However, the model seems to perform poorly on uncertain classes, such as *In the middle, neither yes nor no* with an F-1 score of 26.3%. A potential reason why we achieve a higher performance on certain classes versus uncertain classes could be because the model has not seen enough training data, or because the interpretation of those IndirectQA answers are more vague when scored by human annotators.

[4] `https://github.com/pytorch/fairseq`

4.2.2 T5

Although BERT is a popular choice for pretraining on a large-scale dataset, since its creation, there have been numerous transfer learning frameworks produced that achieve SoTA results on various NLP tasks. One of such frameworks is the “Text-to-Text Transfer Transformer”, also known as T5 (Raffel et al., 2020). In this model[5], the creators adopted a unified approach where the inputs and outputs are texts, allowing for a generalized model to be used in different linguistic tasks. The model is trained on the giant Colossal Clean Crawled Corpus (C4) created by the authors, consisting of 800GB of English text. This model was adapted for our QA task, as the question-answer pair can be passed in as an input, with the label returned directly as a text output. However, this model did not perform as well as the others, as reported in Table 1, possibly due to the ambiguity in the relations between the questions and answers.

4.2.3 UnifiedQA

The UnifiedQA model introduced by Khashabi et al. (2020) aims to learn linguistic reasoning abilities that generalize across input formats. It is trained on eight different datasets used for four different QA task formats. Fine-tuning this model also resulted in SoTA results on 10 QA datasets, so we were interested to see how it would perform on the Circa dataset. We hypothesized that the model’s focus on learning generalizable linguistic reasoning abilities, regardless of input format, would help it to have better performance in understanding implicit meaning.

For our experiments, we used the UnifiedQA-T5-small model checkpoint released by Allen AI[6]. Similar to the baseline experiments, the model was fine-tuned for three epochs on a training set of the Circa dataset, and finally evaluated on the test split. For this model, we did not use any intermediate datasets to fine-tune the model, as UnifiedQA is already pre-trained on eight datasets, including BoolQ which was used in the baseline Circa experiments (Khashabi et al., 2020). The data was pre-processed to fit the UnifiedQA input format requirements. The results of these experiments for both the strict and relaxed settings are shown in Table 1, along with the class-wise F-1 scores.

The model achieved an overall accuracy of 89.6% on the relaxed setting and 74.7% on the strict setting, thus beating the baseline for the relaxed setting, but not for the strict. A closer analysis of the model’s performance using the class-wise F-1 scores shows that the model does not perform well on the classes which have a lower proportion in the data, namely *In the middle, neither yes nor no* for both the strict and relaxed settings, as well as *Probably no* in the strict setting. It is possible that the model did not have enough training instances of these types to learn the linguistic features or indicators for these classes.

5 Results & Conclusion

In order to match the evaluation methods of the original paper, the performance of the transfer learning experiments were evaluated using *Overall Accuracy*, as well as *Total* and *Class-wise F-1 Scores*. The accuracy score measures the rate at which the predicted labels match the original values, and from our experiments the RoBERTa model finetuned on MNLI and Circa had the highest value at 89.6%. As there is an uneven class distribution, we additionally calculated the F-1 score, which is a weighted average of precision and recall. The best performance for F-1 score was achieved by the RoBERTa model in the strict case with 87.5%. Finally, due to the class imbalance within both the strict and relaxed settings, we also calculated the class-wise F-1 scores to gain more insight. The classes with lower frequency scored lower, similar to the original findings on Circa, and require “models to deeply connect the question and answer” (Louis et al., 2020). These results are shown in Table 1, and the class distribution is shown in Table 3 in the Appendix.

The results from our experiments show that although some SoTA models show significant improvement in performance compared to the original baseline of BERT, the dataset is not extensive enough to achieve desired results. Particularly, the class imbalance and the ambiguity of in-between labels pose a challenge to every model. Thus, further development in the corpus is required to achieve advancement in IndirectQA tasks.

6 Collaboration Statement

All team members participated in brainstorming, background research, and writing the report. Each member conducted one of the replication mod-

[5] https://github.com/google-research/text-to-text-transfer-transformer

[6] https://github.com/allenai/unifiedqa

els and the experiments: Anu-Ujin worked on the BERT-YN and T5 models, Angela worked on BERT-MNLI-YN and RoBERTa-MNLI, and Lakshmi worked on BERT-BOOLQ-YN and UnifiedQA.

7 Ethical Considerations

Li et al. (2020) created the UNQOVER framework to study bias through underspecifed questions. In their research, they found that QA models make decisions based on a mixture of reasoning and other stereotypical associations, which they learn from the data they are trained on. Through their experiments using BERT-based models, like RoBERTa and BERT, they found that larger models show more bias and are also prone to positional dependence. Positional dependence implies that model prediction changes based on the order of the subjects, even if the context remains unchanged. Thus, the use of these large language models can itself introduce different classes of stereotypes, such as gender and nationality. Additionally, it is not clear whether the creators of the Circa dataset took any steps to remove biased or stereotypical instances during their data collection process, or if there are cases of underspecified questions with biased labels.

Furthermore, some samples in this dataset require not only linguistic comprehension, but also an understanding of Western culture or practices. For example, one instance in the dataset is as follows:

Question: Do you like to drink?
Answer: I'm in AA.
Label: No

Interpretation of this answer involves not only an understanding of what 'AA' is (Alcoholics Anonymous), but also what that entails about the responder's preference for drinking. Optimizing our models to this data may cause it to have a Western-specific view.

References

Yong-gang Cao, James J. Cimino, John Ely, and Hong Yu. 2010. Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics*, 43(6):962–971.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Nancy Green and Sandra Carberry. 1994. A hybrid reasoning model for indirect answers. *arXiv preprint cmp-lg/9406014*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. *Findings of the Association for Computational Linguistics*.

Stephen C Levinson. 1983. Pragmatics cambridge university press. *Cambridge UK*.

Tao Li, Tushar Khot, Daniel Khashabi, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. *CoRR*, abs/2010.02428.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding Indirect Answers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):167–182.

Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. 2009. Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference*, pages 136–143.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research 21*.

Deborah Rossen-Knill, Beverly Spejewski, Beth Ann Hockey, Stephen Isard, and Matthew Stone. 1997. Yes/No Questions and Answers in the Map Task Corpus. *University of Pennsylvania Institute for Research in Cognitive Science Technical Report*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Xuchen Yao. 2014. *Feature-driven question answering with natural language alignment*. Ph.D. thesis, Johns Hopkins University.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Appendix

Model	**Accuracy for relaxed**		**Accuracy for strict**	
	Original	*Replicated*	*Original*	*Replicated*
BERT-YN	87.8	83.3	84.0	87.3
BERT-BOOLQ-YN	87.1	85.6	83.4	82.1
BERT-MNLI-YN	88.2	86.4	84.8	82.6

Table 2: Replication results in comparison to original values

Class	**Count**	**Proportion**
Relaxed		
Yes	16628	0.496
No	12833	0.383
Yes, subject to some conditions	2583	0.077
In the middle, neither yes nor no	949	0.028
Other	504	0.015
Strict		
Yes	14504	0.460
No	10829	0.344
Yes, subject to some conditions	2583	0.082
Probably yes / sometimes yes	1244	0.039
Probably no	1160	0.037
In the middle, neither yes nor no	638	0.020
Other	504	0.016
I am not sure how X will interpret Y's answer	63	0.002

Table 3: Distribution of Classes in the Full Dataset