Understanding Indirect Answers Using *Circa*

Anu-Ujin Gerelt-Od (ago265), Lakshmi Menon (lsm454), Amber Teng (at2507)

Motivation Question-Answering models are an essential part of NLP research and provide the building blocks of many applications such as automated customer service chatbots and voice assistants. One dataset involved for such models is the *BoolQ* dataset, which was created to test models' inferential abilities for answering naturally occurring boolean questions (Clark et al., 2019), and is now included as a task in the SuperGLUE benchmark for evaluation of natural language models (Wang et al., 2019). However, early research has found that very often, the answer to a yes/no question does not explicitly contain the words 'yes' or 'no', and it 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. To our knowledge, there are currently no other published works implementing this IndirectQA task using the Circa dataset. For this project, we aim to study how we can use existing models and transfer learning techniques to outperform the models suggested in the Circa paper.

Data The Circa dataset consists of 3,431 unique questions with up to 10 indirect answers each, for a total of 34,268 question-answer pairs, and it is publically available through Google Research's Github repository.[1] 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*. The 'strict' scheme has a total of 4 possible classes, while the 'relaxed' scheme has 6 (Louis et al. 2020). A sample question-answer pair and label from the dataset is shown below:

Q: *Do you have the experience to succeed in that job?*
A: *I'm overqualified for the position.*
Label: *Yes*

Modeling and Analysis In the original paper, the IndirectQA task was tested using a BERT model as a baseline, and fine-tuned on the Circa dataset, as well as a combination of BoolQ and MNLI. The best performance resulted from the model fine-tuned 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. We will first try to replicate the paper's results using the same model and analysis methods, and then try to improve the results by using other models in place of standard BERT. Possible models are T5 and RoBERTa, as they are currently leaders in the BoolQ task, as well as UnifiedQA, which could show the impact of pre-training on more diverse question-answer types (Khashabi et al. 2020). Models will be

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

evaluated using overall accuracy as well as class-wise F-scores on a test split of the Circa dataset, following the same setup as the original paper.

Collaboration Statement All team members participated in brainstorming, background research, and writing the proposal.

References

Clark, Christopher, 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, Volume 1 (Long and Short Papers)*

Khashabi, Daniel, 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: EMNLP 2020*

Louis, Annie, 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 (EMNLP)*

Rossen-Knill, Deborah, 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 No. IRCS-97-11.*

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. **Superglue: A stickier benchmark for general-purpose language understanding systems.** *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*