# Making Sense of Commonsense

*How can language models perform better at commonsense reasoning?*

Final Project Proposal
W266 Fall 2020
Sonali Serro, Haerang Lee

## Introduction

*He is beginning work at his first job. What is he feeling? Where in a house would a table be most likely placed? What do parents encourage their children to do when they experience boredom?* Humans find questions like these fairly trivial to answer, but pre-trained language models--even the large ones--often trail behind human accuracy. This highlights the importance of commonsense reasoning in natural language processing, one that parallels our ability to reason about people's behavior and intentions, and our understanding of the physical world.[1]

## Research Question

Commonsense reasoning is considered a challenging NLP task, frequently attributed to reporting bias as common assumptions are often left unspoken. As a result, the goal of our research is to *incorporate additional "common sense" knowledge into the language model in an attempt to improve performance on a commonsense reasoning task.* We would like to experiment with multiple knowledge sources, and subsequently evaluate the performance of each resulting model in comparison to a baseline.

## Dataset

CommonsenseQA[2] is a **multiple-choice question answering dataset** that aims to capture common sense reasoning beyond associations (Talmor et al., 2019). Derived from ConceptNet (Speer et al., 2017), an open multilingual knowledge-base graph, the dataset consists of 12,102 questions, each accompanied by one correct and four distractor answers. The questions are crowd-sourced such that each *question concept* is associated with the multiple *answer concepts* via the same semantic relation (*AtLocation, Causes, CapableOf, etc.)* The authors report that fine-tuning BERT-Large (Devlin et al., 2018) on CommonsenseQA achieves an accuracy of 55.9%, which is substantially lower than human performance, 88.9%. However, recent leaderboard[3] results indicate that this gap is slowly closing.

[1] https://en.wikipedia.org/wiki/Commonsense_reasoning
[2] https://www.tau-nlp.org/commonsenseqa
[3] https://www.tau-nlp.org/csqa-leaderboard

## Proposed Task and Model(s)

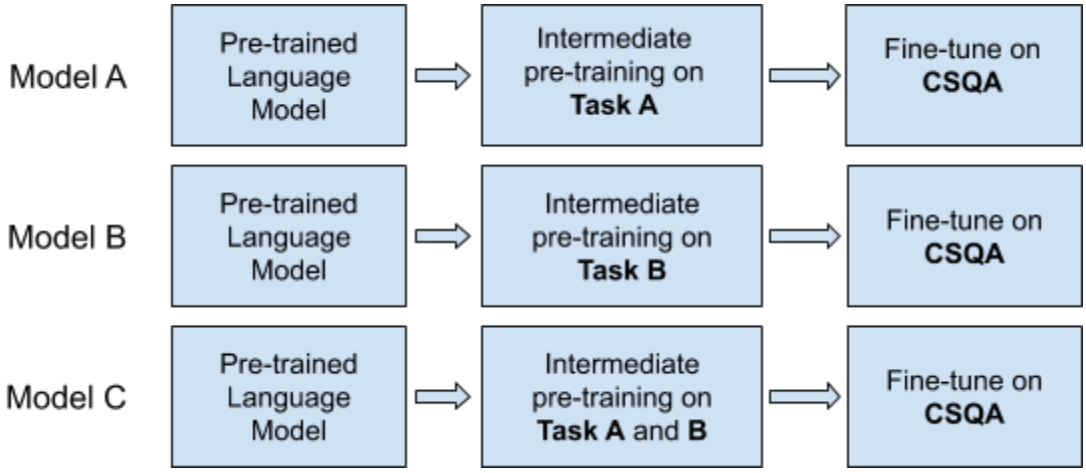
1. **Baseline:** Establish a baseline performance on the CommonsenseQA dataset, using either a BERT or T5 pre-trained language model.
2. **Identify external knowledge source**: Identify 2-3 datasets/tasks to serve as a resource for transfer learning of commonsense knowledge into our model. Here are some potential candidates, given that commonsense encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life (Liu and Singh, 2004).

| Task | Description | Format | Dataset Size (Train, Dev, Test), Average Tokens |
|---|---|---|---|
| **Social IQa** (Allen AI) | Social commonsense intelligence that focuses on reasoning about **people's actions** and their **social implications.** | QA (MCQ) | **Size**: 37K (33K, 2K, 2K)<br><br>**Average Tokens**<br>Contexts: 14.04 words<br>Questions: 6.12<br>Answers: 3.6 |
| **ATOMIC** (Allen AI, U of W) | An atlas of machine commonsense for **everyday events, causes, and effects.** | Knowledge Graph | 877K triples (if-event-then)<br>398K nodes |
| **Cosmos QA** (Allen AI) | Commonsense-based **reading comprehension.** Cause, effect, entity facts, counterfactuals. | QA (MCQ) | 35K (25K, 3K, 7K)<br><br>Paragraph: 69.4 words<br>Question: 10.3<br>Answer: 8.0 |
| **Physical IQa** (Allen AI) | Naive **physics** reasoning that focuses on how we interact with **everyday objects in everyday situations.** | QA (MCQ) | 20K (16K, 2K, 3K)<br><br>Goals: 7.8 words<br>Solutions: 21.3 words<br>3.7M lexical tokens in training |
| **MC-TACO** (U Penn, Allen AI) | **Temporal commonsense** comprehension (duration, temporal ordering, typical time, frequency, stationarity) | QA (MCQ) | 13K unique combo of Q&A<br>2K unique Qs<br><br>Sentence: 17.8 words<br>Question: 8.2<br>Answer: 3.3 |
| **WikiHow** (UCSB) | **Article** and **summary** pairs from WikiHow. Articles are written by ordinary people, not journalists, **describing the steps** of doing a task. | Text Summarization Dataset | 230K article & summary pairs<br><br>Article Length 579.8 words<br>Summary Length: 62.1<br>Vocabulary Size: 556,461 |

**Questions on external knowledge tasks:**

- ➔ Would you recommend any of the above datasets over the others? We want to pick two or three, and the list is in the order of our preference.
- ➔ ATOMIC and WikiHow are not in a QA format. Do you have any concerns about us using these?
- ➔ Should we try our best to keep the dataset size consistent amongst the various knowledge sources to make a meaningful comparison?

3. **Two-Stage Transfer Learning:** Incorporate the additional knowledge into the language model using a Two-Stage Transfer Learning approach. Each model will introduce *additional training objectives by ways of solving related tasks* and will leverage a different knowledge source (or a combination of knowledge sources in a multi-task learning approach.) We will subsequently evaluate each model's performance on the CommonsenseQA test set.
   Below is a diagrammatic representation of the potential architecture.




Time permitting, we would like to explore if the models following the additional pre-training strategy maintain comparable performance on other NLP tasks, potentially on the GLUE benchmark (Wang et al., 2018).

**Questions on architecture:**

- ➔ How important is it to keep the hyperparameters (e.g., number of layers, nodes) consistent across our various pipelines to make a meaningful comparison?
- ➔ We considered multi-task learning (training simultaneously on multiple tasks by mixing the loss functions), but we're concerned about the auxiliary tasks (additional knowledge sources) creating a conflict with the primary task (CommonsenseQA). At most, we might do MTL for two auxiliary tasks (e.g., Model C above). Any thoughts?

## Related Work

- ➔ Does BERT Solve Commonsense Task via Commonsense Knowledge? (Cui et al., 2020)
- ➔ How Additional Knowledge can Improve Natural Language Commonsense Question Answering? (Mitra et al., 2020)
- ➔ Explain Yourself! Leveraging Language Models for Commonsense Reasoning (Rajani et al. 2019)
- ➔ Unsupervised Commonsense Question Answering with Self-Talk (Schwartz et al. 2020)
- ➔ Align, Mask and Select: A Simple Method for Incorporating Commonsense Knowledge into Language Representation Models (Ye et al., 2020)
- ➔ Improving Question Answering by Commonsense-Based Pre-Training (Zhong et al., 2019)
- ➔ Multi-task learning for natural language processing in the 2020s: where are we going? (Worsham and Kalita, 2020)

## References


Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In AAAI, pages 4444–4451.

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

Liu, H., and Singh, P. 2004. Conceptnet a practical commonsense reasoning tool-kit. BT Technology Journal 22:211–226.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.