Social IQA: Social Interaction Question Answering

Sai Srujan Jaligama[1], Bhavya Talluri[1], Nivethan Nadaraj Kumar[1], Akshay Kumar Gunda[1]
[1]Arizona State University
{sjaliga1,srtallur,nnadaraj,agunda1}@asu.edu

Abstract

Question Answering has become a widely researched problem in Natural Language Processing (NLP). Prior benchmarks focused on physical and taxonomic knowledge. This work addresses the Social Interaction Question Answering (Social IQA) problem, the first large scale benchmark for commonsense reasoning about social situations. A major challenge in Social IQA is the matching of answers with respective questions. Question answering is one of the most worked tasks in natural language processing. Most approaches for solving this problem use only the textual content present in training data. This work tackles the question answering with an external knowledge such as a Knowledge Repository by retrieving relevant information from it. We perform our experiments on the Social IQA Dataset (Sap et al., 2019), which has context, question and options.

1. Introduction

Human language offers a unique unconstrained approach to probe through questions and reason through answers about social situations. Social IQA requires a commonsense reasoning for such social situations. The ability to make sense out of the actions of others is critical to people's daily functioning. Humans are social experts, they understand that people's actions are directed at goals and are driven by intentions (Ganaie and Mudasir, 2015). Social intelligence is an aggregated measure of self and social awareness, evolved social beliefs and attitudes, and a capacity and appetite to manage complex social changes, i.e. it is the core nature of the people and how they act in the social system.

2. Related Work

The problem of Social IQA demands common sense in the form of an external knowledge to understand the context and answer the question more efficiently. Hence, we focused on research works which worked on employing external knowledge into the natural language processing tasks and works on common sense question answering.

External knowledge was used for the textual entailment task by using knowledge graphs (Kapanipathi et al., 2019), where a subgraph was chosen using a PageRank algorithm and encoded to capture the semantic information. In another work, answering open domain questions (Chen et al., 2017) was tackled using Wikipedia as a source of knowledge, which uses TF-IDF for document retrieval.

A recent work on common sense question answering (Lin et al., 2019) incorporated knowledge graphs to give machines the ability to make presumptions. By using ConceptNet (Speer et al., 2017) as the external source of knowledge, state of the art performance was achieved. Since we couldn't meet the resource requirements of ConceptNet, we had to go with other knowledge sources.

A new open domain question answering task (Min et al., 2020), introduces ambiguous question answering. The work also introduces a new dataset AMBIQA for training on ambiguous questions. Another work on knowledge base question answering (Chen et al., 2019) proposes to directly model the two-way flow of interactions between questions and knowledge base via their model BAMnet.

3. Dataset/Task Description

Social IQA contains 37,588 multiple choice questions with three answer choices per question. Questions and answers are gathered through three phases of crowdsourcing aimed to collect the context,the question,and set of positive and negative answers (Sap et al., 2019). An example of the Social IQA dataset is as follows:

> **Context**: Bailey was tired of her husband beating her so she filed for divorce today.
> **Question**: how will this make others feel?
> **Options**: **(a)** happy **(b)** downhearted **(c)** frightened

In this example, as Bailey is getting hit by her husband, others have sympathy towards bailey. When she files divorce, others feel happy as she won't get hit by her husband again. This kind of interpretation is easy for humans as they know what marriage is and what does beating mean and what happens when people get divorce as they trivially acquire these social reasoning skills, but this can't be taught to a machine directly. Giving just the premise won't make the machine understand implicit meaning about the premise. So, we need to induce machines with common sense knowledge to make them understand marriage implies connection, beating implies sadness, divorce implies no connection, no connection implies no beating, therefore no beating implies happiness.

4. Methods/ Implementation

We can induce common sense logic into the system by using the following approaches (Pratyay and Baral, 2020)

- Use a pretrained model like BERT, RoBERTa and fine-tune the model with the dataset.
- Induce external knowledge into the model and classify based on that.
- Combine both the methods, i.e. first fine-tune the pretrained model with the dataset and then add external knowledge based on the problem domain into the model.

We need additional external knowledge even though the pretrained models such as BERT, RoBERTa are trained on huge common sense data because most of these models are trained on the data obtained from BookCorpus and English Wikipedia, which may not be sufficient to train the model to possess knowledge of social situations.

4.1 Proposed Approach

External Knowledge can be acquired by following approaches:

- Using the Knowledge Repository: There are many Knowledge Repositories present that are made using different sources. The knowledge repositories are further divided into two types,unstructured and structured knowledge repositories. Examples of some repositories are Atomic, ConceptNet.
- Using the web to scrap the required knowledge: We can also extract the required knowledge from the web using non-stop words and retrieving in a form (for example, in the form of a Knowledge Graph) which will be useful for models to analyse.

4.2 Overview of Fine Tuning BERT with Social IQA dataset:

The Social IQA dataset is in the form of a Json file with context, question, answerA, answerB and answerC as its labels. The input dataset is tokenized to a format with which the BERT was pre trained by using "bert-large-uncased" and then CLS tokens are generated from the formatted input dataset. These CLS tokens are fine-tuned with the BERT model which will be used in classification.

Though BERT does a faster Fine-tuning with the dataset. It's bidirectional approach (MLM) converges slower than left-to-right approaches (because only 15% of words are predicted in each batch) but bidirectional training still outperforms left-to-right training after a small number of pre-training steps. But only using a pre-trained

model for the Social IQA dataset may result in many misclassification errors.

5. Analysis

SocialIQA Dataset consists of positive and negative answers. Negative answers are the options which are generated from the same context but they are correct answers for a different question. These cannot be handled by the regular models as there is a lot of similarity between all the answer choices and regular models cannot differentiate between them. So, we need additional knowledge to handle such problems.

The main misclassifications that occurs when using pre-trained models like BERT ,RoBERTa and ELMo are

- The questions are drawn out of the context.

Context: Bailey sorrowfully confessed to cheating on Jan.
Question: Why did Jan do this?
Options: (a) wanted to humiliate Bailey **(b)** wanted to make amends **(c)** wanted to apologize to Jan
Answer: wanted to humiliate Bailey
Analysis: The context only tells about the action done by bailey. But the question asked about the action done by Jan.

- The incorrect choices are more similar to the context than the correct choice.

Context: Carson liked Cameron enough to ask them to play video games with them.
Question: How would Carson feel afterwards?
Options: (a) Like they want to play games **(b)** a good person **(c)** a friendly person
Answer: a friendly person
Analysis: Context tends to favor option a according to the similarity than option c which is the correct answer.

The following may also cause the misclassification error in BERT:

- The choices are not similar with the context.
- The context has a multiple occurance of a word with multiple meanings (problem due to masking in BERT).

5.1 Analysis of models with external knowledge

We have used 4 knowledge repositories for providing external knowledge to the model; those are Atomic, Wiki, Winograd and our own generated knowledge repository. We have formulated questions which are more similar to the social and emotional situations that we encounter in our daily life and provided different answers along with the reasoning in the generated knowledge repository. As given the same underlying conditions, each social situation can take a different answer, we have used such an approach.

Examples of data from these Knowledge Repositories are as follows:

- Atomic

Alex 'd better go . as a result, Alex wants to resign his job
Alex accepts Pat invitation . as a result, Alex feels good
Alex accepts the fact . as a result, Alex wants to carry on with their work day

- Wiki

How to Make Spaghetti Squash Soup :.: Using a large ripe spaghetti squash wash it carefully and pierce it with a fork
How to Oppose a Motion for Judgment Notwithstanding the Verdict :.: File your documents at the courthouse
How to Answer a Request for Admissions :.: Find your deadline for responding

- Winograd

The treasury workers took the gold bars off of the trolley and stacked them in the safe until the trolley was empty.
Cricket is a favorite pastime of Lindsey, while Rachel has never heard of it. Rachel is more likely to be American.
The man realized he had gotten the lens for his glasses instead of his camera, the camera lens was much thicker.

- Generated

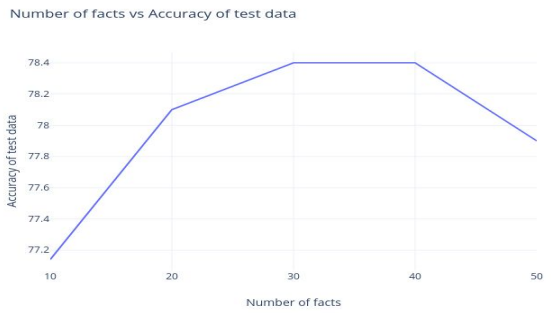
Why do many people have their hands over their mouth . They feel shock and sadness over the situation. Why do many people have their hands over their mouth . They are stopping themselves from saying something. Why do many people have their hands over their mouth . They are shocked by what is being said. Why do many people have their hands over their mouth . They think what is being said is offensive and are shocked by it.

We used part of dev data for validation and remaining for testing the model. The accuracies of the model for 10 epochs are as follows.

Model	**External Knowledge**	**Dev Accuracy**	**Test Accuracy**
Bert-Base-Cased	Atomic	65.10	63.72
Roberta-Base	Atomic	69.52	68.70
Bert-Large (MCQ-Concat)	Atomic	70.50	69.34
Roberta-Large (MCQ-Concat)	Atomic	77.28	75.95
Roberta-Large (MCQ-Weighted Sum) **(baseline)**	Atomic	77.79	77.92
Roberta-Large (MCQ-Concat)	Atomic+ Generated dataset	77.33	74.36
Roberta-Large (MCQ-Concat)	Atomic+ Wiki Data Repository	77.64	76.15
Roberta-Large (MCQ-Concat)	Atomic+ Winograd	77.53	76.82
Roberta-Large (MCQ-Concat)	Atomic+ Winograd	78.33	76.51
Roberta-Large (MCQ-WeightedSum)	Atomic+ Wiki Data Repository	78.84	77.52
Roberta-Large (MCQ-Weighted Sum)	Atomic+ Winograd+Generated Data Repository	77.64	78.66
Roberta-Large (MCQ-WeightedSum) **(final model)**	Atomic+ Winograd	79.00	78.40

The change in accuracies w.r.t the number of facts retrieved using elastic search for RoBERTa - large with winograd dataset is as follows.

Facts	Accuracy
10	77.14
20	78.10
30	78.40
40	78.35
50	77.9



As we can see from the graph as the number of facts retrieved are increased the accuracy increases for a model till a certain value then the accuracy falls as the number of facts increases lot of noise facts are added which will cause the accuracy to fall.

The Common misclassification errors occurred with these models for SocialIQA dataset are:

- Some of the answers are wrong.

Context: Tracy understood Casey's mood better once they sat down and talked it out together.
Question: What will Casey want to do next?
Options: (a) hug Tracy **(b)** reach an understanding with Casey **(c)** be mad at Tracy.
Answer: reach an understanding with Casey
Analysis: As we can see the question asks about Casey and the answer given is "reach an understanding with Casey" rather than "reach an understanding with Tracy".

- The facts are retrieved for each "context + option" pair, so all the three options have the same facts retrieved based on the keywords of the context and differ in the facts retrieved based on the keywords of the options. We are not able to provide enough external knowledge facts to provide good reasoning for the correct option i.e. more related facts between wrong choices and context are being retrieved.
- Some questions ask to describe a person but the correct answer choice is how a person feels.

Context: Skylar got a cast on her leg after she broke it while skateboarding.
Question: How would you describe Skylar?
Options: (a) athletic **(b)** terrible **(c)** sad
Answer: sad
Analysis: The model predicts athletic as that's how Skylar can be described here. Having been trained on questions of "describe" more relevant to a person's character/profession the model predicts athletic with more probability.

- Some twisted questions require external knowledge for every part of the context, the model predicts only considering some part of the context and it's facts and fails to predict the correct answer.

Context: Alex sets up a fund raiser for under privileged children. Alex earns about $5,000 but only gives $2,500 to charity.
Question: How would you describe Alex?
Options: (a) very honorable **(b)** a good person **(c)** untrustworthy
Answer: untrustworthy
Analysis: The model predicts the answer as "very honorable" whereas considering the complete context we can see that the correct answer is "untrustworthy".

- The facts retrieved from these knowledge repositories are not specific to the context of the question, instead out of the many facts retrieved very few are useful for the correct prediction while the rest of the facts are noise data.

5.2 Analysis of baseline with final model:

The baseline for our project is Robert-large with weighted sum and Atomic Knowledge repository using elastic search **,** as the external knowledge using atomic was not sufficient enough to predict correctly. So, we have used different knowledge repositories along with atomic. Out of these Roberta-large weighted sum with atomic and winograd knowledge repositories gave highest accuracy.

Few of the things using winograd which aided to predict correctly in the final model are as follows:

- Winograd provided facts which increased the prediction probability of correct answer choice.

Context: Sasha had a plane to catch, so Sasha left the house within hours of waking up.
Question: Why did Sasha do this?
Options: (a) travel patiently **(b)** fly the plane **(c)** miss the plane
Answer: fly the plane
Winograd fact: She decided to leave coach and make her way to catch a plane so she could get there faster but the plane actually got there slower due to delays.
Analysis: Winograd has given us this fact which is vital to answering this question. The fact related to this context and question is not present in Atomic.

- Winograd provides related facts of both the context and option pair completely which helps

to answer twisted questions where other models failed to do so.

Final model failed to answer questions where the correct option varies based on the perspective of the person.

Context: Alex couldn't believe they were able to make the escape to Mexico easily.
Question: What will Alex want to do next?
Options: (a) celebrate **(b)** avoid being caught after committing their crime **(c)** get arrested
Answers: avoid being caught after committing their crime
Analysis: Our model predicts "celebrate" as the answer, looking in one of the perspectives having escaped so narrowly people tend to celebrate such situations whereas the correct answer choice " avoid being caught after committing their crime " is also true.

6. Contributions:

In the given baseline usage of external knowledge along with the Roberta-large model eliminating noisy data using reranking procedure was the key. So, We have used different knowledge repositories like wiki, winograd and our own knowledge repository along with atomic which provided more related information about social and emotional situations than using only atomic whose training and testing accuracies has been mentioned above.

We have analysed what type of questions do these models fail like twisted questions, close answer choices, questions which can be seen in multiple perspectives yield different answers and questions which ask to describe a person or their feelings.

7. Conclusion and Future Directions

Social IQA is a common sense reasoning task for social situations. We have used a pretrained model of RoBERTa along with knowledge repositories for the classification, but using only this model is not completely efficient.

We need additional external knowledge even though these models are able to generate nearly 80% of dev accuracy which are specific to the required information. Training on huge common sense data of social situations is not sufficient. We have to deduce a mechanism to eliminate unrelated information for a particular context more efficiently along with having a mechanism to eliminate choices which even though may be correct, correlated to the context but in the current situation is not the best answer which is very difficult to obtain for a machine.

References

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, Yejin Cho. SOCIALIQA*: Commonsense Reasoning about Social Interactions*. In EMNLP, 2019.

MY Ganaie and Hafiz Mudasir. 2015. *A Study of Social Intelligence & Academic Achievement of College Students* of District Srinagar, J&K, India. Journal of American Science, 11(3):23–27.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral.*Natural Language QA Approaches using Reasoning with External Knowledge*. In arXiv March 2020, *arXiv:2003.03446v1*.

Danqi Chen, Adam Fisch, Jason Weston & Antoine Bordes, *Reading Wikipedia to Answer Open-Domain Questions*, In arXiv:1704.00051v2 [cs.CL], 28 Apr 2017.

Pavan Kapanipathi, Veronika Thost, Siva Sankalp Patel, Spencer Whitehead, Ibrahim Abdelaziz, Avinash Balakrishnan, Maria Chang, Kshitij Fadnis, ChulakaGunasekara, Bassem Makni, Nicholas Mattei, Kartik Talamadupula, Achille Fokoue, *Infusing Knowledge into the Textual Entailment TaskUsing Graph Convolutional Networks,* In AAAI, 2020.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. "*ConceptNet 5.5: An Open Multilingual Graph of General Knowledge*." In proceedings of *AAAI* 31.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, Xiang Ren, *KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning*, In arXiv:1909.02151v1 [cs.CL], 4 Sep 2019.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, Luke Zettlemoyer, *AMBIGQA: Answering Ambiguous Open-domain Questions,* In arXiv:2004.10645v1 [cs.CL], 22 Apr 2020.

Yu Chen, Lingfei Wu, Mohammed J. Zaki, *Bidirectional Attentive Memory Networks for Question Answering over Knowledge Bases*, In arXiv:1903.02188v3 [cs.CL], 28 May 2019.