Distribution Agreement

Daniil V. Huryn   March 2, 2022

Framework for Automatic Generation of Large-scale Dialogue Data from Online Forums

By

Daniil V. Huryn

Jinho D. Choi
Adviser

Computer Science

Jinho D. Choi
Adviser

Jonathan Hulgan
Committee Member

Ting Li
Committee Member

Framework for Automatic Generation of Large-scale Dialogue Data from Online Forums

By

Daniil V. Huryn

Jinho D. Choi
Adviser

An abstract of
a thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Computer Science

2022

Abstract

Framework for Automatic Generation of Large-scale Dialogue Data from Online Forums

By Daniil V. Huryn

Unsupervised Machine Learning models have taken the Natural Language Processing world by storm. Transformers, the currently most popular unsupervised models, utilize vast amounts of data and deliver performance far beyond what could have been achieved only a few years ago. As good as these models are, they have one major requirement - a lot of data. One of the first transformers, BERT, was trained on 3.3 Billion words of data, and later models have used even more (GPT-3). This presents unsupervised dialogue models with a bit of a problem: there's not that much high quality dialogue data out there, certainly not on the scale required. Because Dialogue is far harder to find online then posts, articles, etc., high quality datasets are usually very limited in size (Switchboard, Daily Dialog), while high quantity datasets (Opensubtitles, Reddit Corpus) are either of extremely low quality or of a very specific type, for instance movie subtitles. One of the main mitigations of this issue has been to first train models on large amounts of low quality data, and then fine-tune on low amounts of high quality data. In this paper, we propose a different solution: to create a high quantity, medium quality, multi-turn dataset, that will allow for far better model training. To do this, we intend to utilize a more computational approach to dialogue creation, where we create it from a set of Reddit posts and their respective comments, blending it in a way that creates a new conversation out of a disjointed online forum post. By utilizing the nature of Reddit threads and a variety of Natural Language Processing metrics, we intend to first construct and then thoroughly filter conversations to automatically create a large dataset of high quality dialogues.

Framework for Automatic Generation of Large-scale Dialogue Data from Online Forums

By

Daniil V. Huryn

Jinho D. Choi
Adviser

A thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Computer Science

2022

Acknowledgments

I am very grateful to my advisor Dr. Jinho D. Choi. Dr. Choi helped develop my passion for NLP, and gave me an opportunity to do research in the Emory NLP Lab, something I did not think I would be able to do due to my being from the Oxford Campus. Dr. Choi has given me countless learning opportunities, and taught me much of what I know about research.

I am grateful to Dr. Jonathan Hulgan and Dr. Ting Li. I took many classes with Dr. Hulgan, and the math focused computer science course really helped me understand the foundations of computer science. Dr. Ting Li was my Computer Science professor at Oxford, and I am thankful to her for giving me more of a passion for it.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Dialogue Datasets

Dialogue Datasets are extremely important in NLP research. As varied as current machine learning approaches are, they all have one thing in common: they need data to train on. The most common application of dialogue datasets is training chatbot models, like Blenderbot [8] and DialoGPT [10]. Dialogue is an incredibly complicated thing. In, say, a blog post, only one person is developing a train of thought. In a dialogue, there are two people interacting, where not only do both of them have their own thought process, knowledge, and motivation, but they also react to what the other person says. Even worse, sometimes people do not need to react to the other person, or it is expected that they react in a very specific way (for instance, there are many times when making a joke is improper). Not only does the meaning of a persons reply has to be perfectly adapted to the previous statement, but often there are other requirements, where saying something like ”Anyway, ...” might be considered a rude disregard for the other person's statements. Dialogue having two alternating points of view makes understanding it much more complicated for neural models, and, while more advanced models will understand it better, one can't get around the fact that in the world of machine learning, more data is one of the best ways of improving performance of practically any model, provided it is complex enough.

Dialogue is not only more complicated than most other text sources, it is also harder to find. While most blogs, books, speeches, etc. can be found online, most people keep their dialogues private, either in person or through a messaging app, making the creation of such a dataset a much more involved task. Currently, there is a large split between low quality ”scraped” data, data that has been automatically collected from online sources, and high quality human generated data, data that has been generated in an extremely controlled environment, taking up far more resources. We will talk more about current popular dialogue datasets in the Background section.

1.2 Intentions

At the moment, there is little research done on computationally manipulating data, rather data is collected and then heavily filtered. A more complicated approach would be to have a large amount of data, and then ”construct” a dialogue out of that data, using components of the original data to build a new dialogue. This is naturally hard to do, as dialogue requires coherency and consistency, limiting the amount of compatible data, as well as requiring sentences to ”flow” properly from one to another (a greeting is usually followed by a greeting, questions need answers, etc). On the other hand, this approach would have both the scalability of scraping dialogue, along with the quality of hand crafted dialogue, since we have more flexibility in our dialogues than scraping, where you take what you can get.

A solution to this problem (and indeed the solution we utilize in our approach), is to use the nature of online data. For instance, comments on a forum post will be mostly limited to the same topics as the original post, and often commenters will respond to specific questions posed by the post, or respond to points it makes. This means that, if one could match corresponding sections, its possible to construct a dialogue of better quality than just random comments, but in a larger quantity than human crafted data. Reddit also has comment threads (people who respond to other comments, creating a comment chain), which allows for us to utilize threads as little pieces of connected dialogue, which are typically far more likely to be related than two arbitrary comments from a post.

Not only would this allow us to create an enormous amount of dialogues (Reddit has over 300 million posts every year), but a key advantage is that we can create dialogues focused on specific topics. By using Reddit posts from ”r/books”, a community focused on discussing books, we can create a dataset of dialogues about books, which might be very useful if one was making a chatbot for a library's website for example.

1.3 Thesis Statement

In this paper, we intend to make a model that will automatically create a high quantity, high quality, multi-turn dataset, that will allow for far better model training. To do this, we intend to utilize a more computational approach to dialogue creation, where we create it from a set of Reddit posts and their respective comments, blending it in a way that creates a new conversation out of a disjointed online forum post. By utilizing the nature of Reddit threads and a variety of Natural Language Processing metrics, we intend to first construct and then thoroughly filter conversations to automatically create a large dataset of high quality dialogues.

Chapter 2

Background

2.1 Current Datasets

There is a clear trade-off evident in current datasets: quality or quantity. Datasets prioritizing quality have a very low quantity (Switchboard [4], DailyDialogue [7]), since they need to have humans interact in a natural way, with manual filtering processes that ensure conversation quality. Datasets with a large number of conversations are usually automatically scraped from the internet, most often from forums (Reddit, Twitter). Due to the nature of online discourse, these are often low quality, and typically barely resemble human-like conversation. Modern improvements to datasets often improve additional annotations to data, such as ”persona” information for speakers (PersonaChat [9]), but naturally this skews even more into the quality side of quality over quantity. Some datasets try to lean into having a lot of domain specific topics, like MultiWOZ ([1]).

Even datasets of supposedly high quality can vary extremely. Switchboard is well known for having very good dialogues, but only a small amount of them (2400). Some papers even release heavily annotated datasets of only a few hundred dialogues. On the other hand, Topical-Chat [5] has human generated dialogues, and as part of the dialogue creation humans were given a set of facts to work with. Unfortunately, this lead to many people simply reading off of the facts, and creating many extremely similar conversations, like seen in Figure 2.1. As you can see, they contain a significant amount of exact repetition, with even more approximate repetition. The data is still useful, but this needs to be kept in mind when using it. This dataset has a trade-off to get this big, and it is not alone - most larger datasets have some drawbacks.

Conversation One	Conversation Two	Conversation Three
Speaker 1: ”do you like shakespeare?” Speaker 2: ”I do. Did you know that he popularized many words and phrases?” Speaker 1: ”Yes like, good riddance in my heart of hearts and many more” Speaker 2: ”Yes. He lived at the same time as Pocahontas too” Speaker 1: ”I wonder what it might have been like had they met somehow”	Speaker 1: ”Hello do you like Shakespeare's work?” Speaker 2: ”I love his work actually. He even popularized many words and phrases we use today” Speaker 1: ”Yes like, good riddance, in my heart of hearts, wild goose chase. He was amazing” Speaker 2: ”Yeah. He lived at the same time as Pocahontas too” Speaker 1: ”Yeah they only dies a year apart, small world” Speaker 2: ”Yeah it is, I wonder what play he would have written had he known about her”	Speaker 1: ”Nice. Do you like Shakespeare?” Speaker 2: ”Yes I do. Do you know that he popularized many phrases” Speaker 1: ”Yes like good riddance, in my heart of hearts and such” Speaker 2: ”Yes and then he also invented names like Jessica, Olivia and Miranda” Speaker 1: ”Yes. And for his works you have to use old english for it to make sense” Speaker 2: ”Yes otherwise the rhymes and puns do not seem to work out” Speaker 1: ”Yes. He lived at the same time as Pocahontas too” Speaker 2: ”I wonder if they met how that would go from there”

Table 2.1: Three separate Topical-Chat conversations. Note that much of it is the exact same, and much is extremely similar.

One of the common approaches to the quality/quantity trade-off is to first train on a large amount of low quality data, and to then further fine-tune the model on high quality data, or even specific types of high quality data to enhance aspects of the model, such as empathetic responses. For instance, Blenderbot trains on 1.5B of reddit data, and then fine-tunes on Empathetic Dialogues [8], BST [8], and others. This approach has lead to some of the best current chatbots, and is the current status quo.

See Table 2.2 for an overview of popular dialogue datasets and their sizes.

Table 2.2: Popular Dialogue Datasets

Dataset	Number of Utterances
Switchboard	200k
Persona-Chat	162k
Topical-Chat	188k
MultiWOZ	116k
BST	76k
ConvAI2	140k
Wizard of Wikipedia	194k
Empathetic Dialogues	50k

2.2 Models

2.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) [2] is one of the first NLP Transformer Models. It had results that surpassed previous ones by far, and started a new trend of transformer models. One of the tasks it was trained on was Next-Sentence Prediction, a task that was dropped in later transformer models. This task was simple: given two sentences, does the second one come after the first sentence in the text that contained the first sentence, or is it unrelated? In other words, is the first sentence the context of the second? We heavily utilize this predictive ability in our approach. We use BERT-Large-Uncased, which has 336 million parameters.

2.2.2 Blenderbot

BlenderBot is an advanced open-domain chatbot. It is meant to talk on any topic, and is trained on a large amount Reddit data and then fine-tuned on three tasks: Blended Skill Talk, Wizard of Wikipedia, and Empathetic Dialogues. It is one of the best current chatbots, and is very good at varied topics, making it perfect for our work. Our approach utilizes it to smooth other disjoint part of our dialogue.

2.2.3 GRADE

GRADE (Graph-enhanced Representations for Automatic Dialogue Evaluation) [6] is an automatic metric we utilize for post-filtering our conversations, and is adept at finding the most incoherent ones. GRADE uses both local contextualized information (using BERT) about how the sentence fits in the dialogue, and also topic-level graph information to look at whether the sentence matches the topic being discussed. It then feeds these two branches into a single predictive model.

Chapter 3

Approach

3.1 Data

Our main data source is Reddit. We collect Reddit posts and their respective comment threads to construct conversations. A typical Reddit post usually has some content in the main post (we looked at posts with textual content, i.e. no links, pictures, etc.). An example can be seen in figure 3.2. Note the body of text in the original post, and the hierarchical comments below it. Top level comments typically respond directly to points the post is making, while lower level comments will often have a mix of response to the post and to a higher comment.

Initially, we also considered Twitter, as it tends to have slightly more natural language in our experience, with fewer forum-specific quirks and abbreviations (like TL;DR - Too Long Don't Read). However, we found that Twitter had two main deficiencies compared to Reddit. First, it has a post limit much smaller than Reddit, only allowing 280 characters, severely limiting the length of our conversations. Second, Twitter tends to have much fewer deep threads, while on Reddit deep comment threads are the norm, rather than an exception.

Because we are utilizing a single post and all of its comments to create a conver-

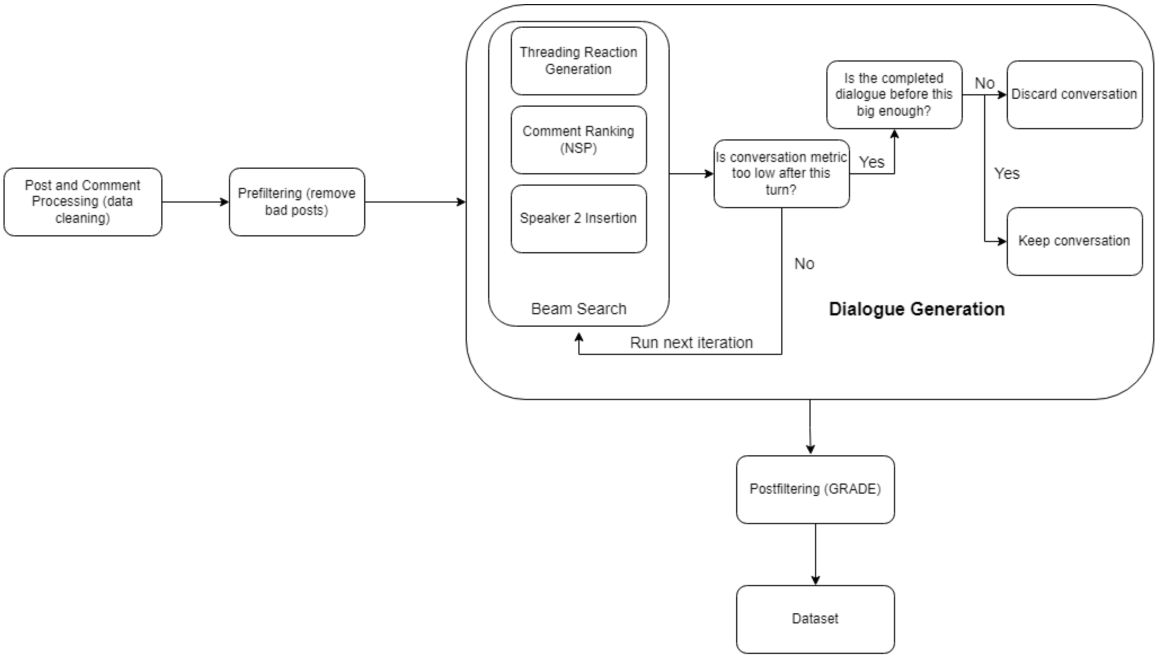
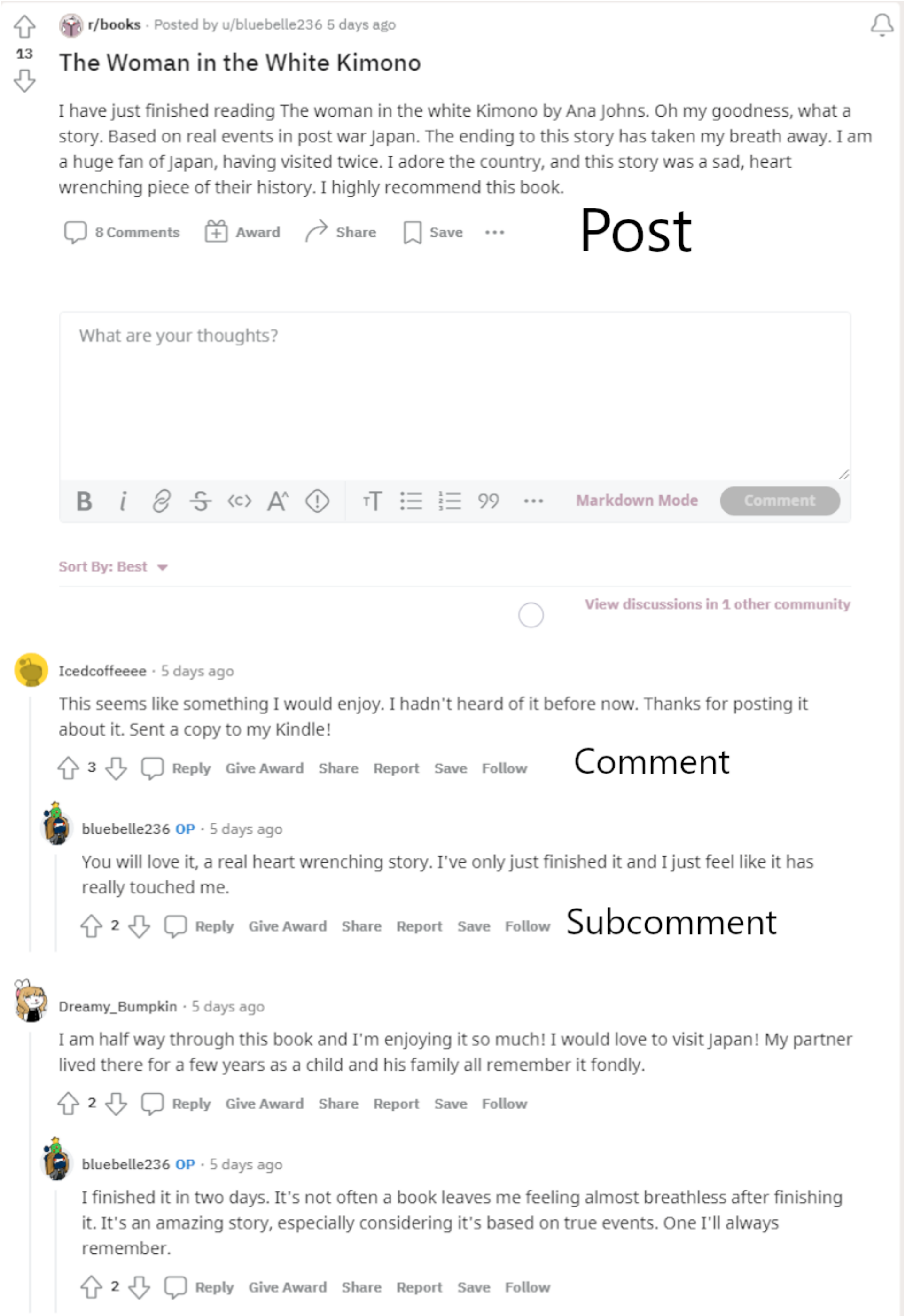

Post and Comment Processing (data cleaning)
Prefiltering (remove bad posts)
Threading Reaction Generation
Comment Ranking (NSP)
Speaker 2 Insertion
Beam Search
Is conversation metric too low after this turn?
Yes
No
Is the completed dialogue before this big enough?
No
Yes
Discard conversation
Keep conversation
Dialogue Generation
Run next iteration
Postfiltering (GRADE)
Dataset

Figure 3.1: Main Architecture of Approach

sation, we can scrape specific "subreddits" that typically focus on some topic/area of interest. This allows us to create dialogues focused on the topics we want, and for this project we looked at: College-related subreddits, Movie-related, and Book-related. The ability to do so is actually one of the key strengths of our approach, as we can generate data for an arbitrary topic (almost anything you can think has a subreddit for it, where there have been countless discussions of every detail of the subject matter). So not only is our data automatically generated, we can specifically target tasks that require certain topics of dialogue.

Processing the data was extremely important, as we are taking raw data off of the internet. We needed to filter out text that contained links, specific Reddit peculiarities (like people putting a "[SPOILER] tag on sensitive content), and inappropriate content (vulgar text). We also had to heavily process the actual text, as Reddit would often contain empty spaces and rare characters for formatting that our model would not be able not understand. Overall, our preprocessing filtering process looked like:

1. Check if it is in English



The Woman in the White Kimono

I have just finished reading The woman in the white Kimono by Ana Johns. Oh my goodness, what a story. Based on real events in post war Japan. The ending to this story has taken my breath away. I am a huge fan of Japan, having visited twice. I adore the country, and this story was a sad, heart wrenching piece of their history. I highly recommend this book.

Icedcoffeeee · 5 days ago

This seems like something I would enjoy. I hadn't heard of it before now. Thanks for posting it about it. Sent a copy to my Kindle!

3

bluebelle236 OP · 5 days ago

You will love it, a real heart wrenching story. I've only just finished it and I just feel like it has really touched me.

2

Dreamy_Bumpkin · 5 days ago

I am half way through this book and I'm enjoying it so much! I would love to visit Japan! My partner lived there for a few years as a child and his family all remember it fondly.

2

bluebelle236 OP · 5 days ago

I finished it in two days. It's not often a book leaves me feeling almost breathless after finishing it. It's an amazing story, especially considering it's based on true events. One I'll always remember.

2

Figure 3.2: An example Reddit post. At the top is the post text, and below are all the comments, along with comment threads (people replying to comments)

2. Check if it is a Reddit peculiar tag like [DELETED]
3. Check if it has invalid characters
4. Check if it has a url
5. Check if it has vulgar/inappropriate content
6. Remove formatting characters, like the empty space character
7. Check if length is too short too use
8. Check if post is only a title or only one sentence

3.2 NSP Approach

3.2.1 Initial Approach

We started off with a simple idea: use BERT NSP (Next-Sentence Prediction) to insert comments along the post. We had a few simple approaches while developing this, and while they were not that good, they were useful in developing our later approach, both in showing which tools worked well (BERT NSP), and saving some of the coding work later. One of the first thing we tried was to simply take the Reddit post, and have a single comment chosen by NSP after it. While the conversations produced were of passable quality, we found this approach to be insufficient for two main reasons: 1) It only produced two-turn conversations, and 2) the post was almost always much longer than the comment, containing many points, with the comment containing one or two points, effectively ignoring the rest of the post. A similar approach was to simply take random comment pairs, and using NSP to choose the best pairs, keeping those as two-turn dialogues. This would provide a higher quality alternative to usual Reddit datasets, which often just take random comment pairs. This approach showed us just how effective NSP was in our preliminary tests, and was a great marker for us of where to go next, but could not supply multi-turn conversations.

Our initial working approach would simply go through the text of the post, and after every sentence check from a list of comments, which sentence fit in more at that spot. This would be a reply of a ”Speaker 2”, where the Original Post would become ”Speaker 1”. After that we would continue, and after every sentence from the original post, insert another comment. Naturally, this simplistic approach did not perform the best, but the results were surprisingly good for the low complexity of the approach. The initial approach algorithm can be seen in Algorithm 1.

After that, we made a couple of simple improvements to this approach. First, we decided to not always put a reply in, as we might want to have several sentences from the original post in a row. Next, we decided stop the approach if no comments under a certain NSP threshold were found.

Our main decision metric was BERT NSP, which, as mentioned earlier, is specifically trained for NSP, with its NSP head predicting whether a sentence should follow another one. In most of our later approaches, we used the entire constructed conversation up to a point as the context (left hand term in NSP(A,B)).

For an example of a dialogue generated by this approach, see Figure 3.3. While most of the utterances do make some sense in relation to the previous one, there is still a significant disconnect at times, and some of the transitions from Speaker 2 to Speaker 1 are especially rough.

Generated from Post about PhD

Speaker 1: ”Hello! I am interested in pursuing a PhD related to climate change and urban development”
Speaker 2: ”Even though I have some contact with the area of urbanistics and climate change, I can’t say whether the PhD would be beneficial without knowing a lot more specifics about your career goals and background”
Speaker 1: ”One seemingly age old question is what the job prospects are after pursuing a PhD - that one should not assume just because they have a PhD, their job prospects are better”
Speaker 2: ”If you are into sustainable urbanism you don’t necessarily need a phd, maybe a Master with an applied focus could be of more interest for your career goal? I don’t know what is the angle you are into but I’m thinking you could check out geography, urbanism and public management departments and see the projects going on in the labs and research centers there”
Speaker 1: ”Additionally, I am not necessarily interested in staying in academia after obtaining a PhD”
Speaker 2: ”For some career goals, a PhD would be to your detriment”
Speaker 1: ”I would like to put my gained knowledge into practice in some other area, such as the private sector, NGOs, etc. I am wondering if anyone has recommended resources for learning more about job prospects given a specific PhD? I think this would be useful to any person currently thinking of pursuing a PhD program”
Speaker 2: ”Ideally it would be to talk to people in this field, both at university and on the market to know what a PhD could give you/ limit you for your career goals”

Figure 3.3: Base NSP Example

```
Algorithm 1: Base Approach
Input   : Post Text, Comments
Output: Dialogue
repeat
    Take next post sentence;
    Rate each comment from comment pool, take best and delete it from pool;
    // We rate using conversation unto that point as context;
    Insert highest rated option into conversation as other speaker;
until End of Post;
```

3.2.2 Early Improvements

Naturally, our approach had promise but needed a lot of improvements for our desired level of performance.

Better NSP Utilization

In the Base approach, we simply looked at $NSP(SentencefromPost, Comment)$ and chose the best comment. This did not consider how the next sentence of the post would fit after the comment, so we switched to:

$$0.5*NSP(SentencefromPost, Comment)+0.5*NSP(Comment, NextSentencefromPost)$$

See Figure 3.4 for an example. With the old NSP approach, you can see that "Fantasy is an amazing genre" is said by both Speaker 1 and Speaker 2, this is because the NSP did not account for the next sentence from the original post, and therefore could not rank it lower due to the repetition. The Better NSP approach accounts for this, and the conversation seems much more natural thanks to it.

Old NSP Conversation	Better NSP Conversation
Speaker 1: ”So, I’ve been thinking about reading Fantasy.” Speaker 2: ”You should absolutely do that, fantasy is an amazing genre!” Speaker 1: ”Fantasy is an amazing genre, and I think I should get into it.”	Speaker 1: ”So, I’ve been thinking about reading Fantasy.” Speaker 2: ”That’s an amazing idea!” Speaker 1: ”Fantasy is an amazing genre, and I think I should get into it.”

Figure 3.4: Example of how Better NSP improves conversations

Better Comments

Initially, we had simply been considering whole comments. However, as posts are large, comments often respond to multiple sections of the post separately. To account for this, we used individual sentences of the comment. Later, we started using sequences of comment sentences up to 3 sentences long (higher was too computationally expensive). In other words, if a comment consisted of sentences A B C, we would consider: A, B, C, AB, ABC, BC.

In Figure 3.5, you can see an example of comment segmentation improving the conversation. The original comment was ”Oh, for sure. I think its the pure fun aspect of it, its not meant to be realistic. Also, I disagree with your second point.” It was awkward to place it in with the old comment code, because while the comment agreed with the first sentence of the original post, and the last sentence of the comment referenced a later part of the post, which seems out of place in the conversation. Better Comments allowed us to cut out the last sentence and consider the first two of the comment, which ended up working perfectly in this example.

For a more detailed explanation of how we segment our comments, see Algorithm 2.

Old NSP Conversation	Better NSP Conversation
Speaker 1: ”Mario Kart has gotta be the best multiplayer game of all time.” Speaker 2: ”Oh, for sure. I think its the pure fun aspect of it, its not meant to be realistic. Also, I disagree with your second point.” Speaker 1: ”And the powerups are so fun too!”	Speaker 1: ”Mario Kart has gotta be the best multiplayer game of all time.” Speaker 2: ”Oh, for sure. I think its the pure fun aspect of it, its not meant to be realistic.” Speaker 1: ”And the powerups are so fun too!”

Figure 3.5: Example of how Better Comments improves conversations

Algorithm 2: Comment Segmentation

```
Input  : Post Top-Level Comments
Output: Varied-size Comment Segments
comments ← set to comments in post;
final_comments ← set to empty list;
repeat
    comment_sentence_list ← list of comment's sentences;
    counter ← set to size 1;
    repeat
        segment_size ← set to 1;
        repeat
            Append entries counter through (counter + segment_size) to
             comment_sentence_list segment_size+ = 1
        until Segment Size is 3;
        counter+ = 1
    until Counter reaches the length of the comment sentence list;
until No comments left;
```

3.2.3 Beam Search

In our Base Approach we simply chose the best option at each point and continued on from then. A significant improvement was adding Beam Search. Instead of choosing the best option, Beam Search chooses the n best options at each point according to our metric (NSP). It then continues along n paths, and after a step along each path we again take the n best options (from among all the paths). See 3.8 for an example. If we simply have one beam, then we select the best option at each point as before, i.e. greedy search. A more formulaic description of our beam search implementation can be seen in Algorithm 3.

In the example, we see that the lack of beam search is very limiting, as a given comment might fit in much better later, but without beam search the metric simply cannot account for this. This example also helps illustrate something that we thought would be a great improvement to the model. We'll expand more on this in the Experiments section, but in our opinion, beam search was one of the most significant improvements to our model, along with Threading and GRADE.

Algorithm 3: Beam Search with size n

```
Input  : Post Text, Comments
Output: Dialogue
beams ← set to list with one empty beam;
repeat
    options ← empty set;
    repeat
        Append generated options from this beam to options
    until End of Beams;
    From all options, choose n best ones.;
    Create new beams list using best options and the beams they were
      selected from.;
until End of Post;
```

3.2.4 BlenderBot

We utilized BlenderBot 2.0 to smooth over rough parts of our conversation. When selecting the best response to a comment from the original post, we also generated a Blender response and considered it as an option. A later improvement was to also use Blender to smooth the transition back to the original post. This was a pain point of our approach, for example: if the original post has sentences A B C, and a comment D fits perfectly after sentence A, but does not fit before B, we either prepend a blender response to B, or append it to D, considering A $\rightarrow D + Blender \rightarrow B; A \rightarrow D \rightarrow Blender + B$.

Blender was given the previous context of the whole conversation. It should be noted that when testing Blender, we needed a significant improvement, as it was by far the most computationally expensive part of our approach, taking 3x longer than all other parts of the approach combined.

Incorporating Blender into our approach was one of the longest parts of our project, as the Facebook package it is part of, ParlAI, did not have an easy way of using Blender without using the entire codebase due to their structure of libraries and dependencies. In the end, we had to run BlenderBot 2.0 as a sub-process that we piped into and out of, which also required a lot of work and debugging due to inconsistencies in how it accepted standard input (it would only read in parts of the line at a time).

3.2.5 GRADE

We use GRADE to ensure coherency in our dialogue. BERT NSP is good at predicting the next sentence given the past context, but it cant fully account for context, especially in longer conversations. GRADE as a metric is not that good at finding the best conversations, but we found it extremely useful in identifying the conversations with the least coherency. We used GRADE to remove the lowest scoring conversations, ensuring a base level of coherency.

An example of a post filtered out by GRADE can be seen in Figure 3.6. This conversation was the lowest rated by GRADE in a batch of 100 conversations, and the utterances are disconnected from each other, for instance the very first utterance seems like it relies on non-existing context, and the second utterance likewise seems entirely disjoint from the rest of the dialogue. Our initial observations confirmed that GRADE would remove the dialogues we considered the worst, and as seen later on in the Experiments section, this was backed up by our evaluation.

3.2.6 Fine-Tuning BERT

One of the core advantages of Transformer models is knowledge transfer [2]. They are first trained on large amounts of general data. and then fine-tuned on particular tasks. That's why we fine-tuned the NSP head on dialogue data, to adapt it more

Generated from Post about Financial Aid

Speaker 1: ”Just recently found this and boy am I salty”
Speaker 2: ”I lived with my sister and her wife from 2017 until I graduated in 2018, and up until I moved out last year, but they worked at a factory at the time and made money so I wouldn’t qualify, despite them literally not being prepared for me to be dumped on them and thus not having money saved or contributing to my education”
Speaker 1: ”Me: *emails school to say I’m doing the required forms and stuff, asking what we’re doing for my FAFSA since I couldn’t complete it due to my mom being garbage and me going no contact with her and stuff, was looking for some sort of independent thing so I would be able to use my own taxes or something instead of hers, counselor had previously said we’d do it when I had been accepted and everything*Counselor: *forwards our email to a financial person*Financial person: *says the best thing for me to do is fill out the FAFSA and we’ll go from there*”
Speaker 2: ”Yep”
Speaker 1: ”I’m livid that these people can’t read”
Speaker 2: ”Soon”

Figure 3.6: Example of post filtered out by GRADE

towards our specific task. We used Facebook’s Multi-Session Chat dataset to fine-tune it. This was a minor improvement, but a necessary experiment, as fine-tuning Transformers is standard practice.

3.2.7 Reactions/Threading

Reactions were also used to smooth over transitions. Just like the Blender Response earlier, we prepended them to the next sentence from the original post, after the inserted reply. Instead of using Blender to generate a response, we looked at a list of pre-recorded reactions like ”Oh wow!” and ”I’m sorry to hear that.”, from which we chose the best scoring option (or none at all if none performed better than default). A later improvement we made on this was to instead look at replies from short comments related to the comment we are using, as we know those are closely related. We called this Threading.

An example of Threading can be seen in Figure 3.7. ”Agreed and I feel the same way” is the comment sentence that Threading prepended to the original post (Speaker 1). As you can see, it improves how Speaker 1 seems to react to Speaker 2, acknowledging his statement more. We found that this was one of the most significant improvements we made, in part due to the large variety of comments available on every Reddit post, and this allowed us to utilize that variety to make the algortihm much more flexible.

3.2.8 QA Model

We considered utilizing a Question-Answering Model to help answer Reddit posts that had questions in them, helping form a cohesive dialogue. For instance, if at some point a user said ”What year did X come out in again?”, our model would help answer that, something NSP would not be able to do (as it can only tell if a response seems appropriate, not whether it is factually accurate). However, based on our testing SotA (State-of-the-Art) Question-Answer models simply cannot handle the variety of questions found in Reddit data, simply due to 1) many questions requiring very specific factual knowledge, which is often not easily searchable (or else a user would not ask on a forum), and 2) their being phrased extremely informally and using reddit slang, like ”DAEF” - ”Does anyone else feel”. Due to its poor performance in these situations (and these situations occur very frequently in Reddit data), we did not include the QA model in our final approach.

3.3 Alexa Grand Prize

Last year, we took part in the Alexa Grand Prize competition, in the team from Emory University. It is a competition where 8 chatbots from top university teams compete for ratings, first with Alexa users, and later with Amazon selected judges. As

Pre Threading Conversation	Threading Conversation
Speaker 1: ”Having read Oliver Twist, I realized that each of the villains represent one of the three types of well-written villains Nancy and Charlie Bates represent the villains you felt sorry for, because unlike the rest of the guys, they actually have empathy:. Nancy tries to save Oliver from the crime life and Charlie Bates was horrified by Nancy's death and even exposed Bill before changing his ways for good.” Speaker 2: ”Bill Sykes, on the other hand, was a villain I very much hated.” Speaker 1: ”They both represented redemption.. Fagin and Mr. Bumble were the villains you hate with a burning passion.. Fagin was a jerk who forced children to work as criminals without any care of how they feel, while Mr. Bumble was a fat hypocrite who saw the people of lower class as non-humans who deserved to be treated poorly.” Speaker 2: ”I didn't hate Fagin because even though he was a shameless crook who used kids to do his dirty work, he only partook in a victimless crime, abhorred violence, and didn't want to see the kids get hurt.” Speaker 1: ”And finally, Bill Sikes and Monks were the villains who were too epic to hate.” Speaker 2: ”Putting children in danger for your own gain without care for their well-being and threatening to murder them isn't epic in my book.”	Speaker 1: ”Having read Oliver Twist, I realized that each of the villains represent one of the three types of well-written villains Nancy and Charlie Bates represent the villains you felt sorry for, because unlike the rest of the guys, they actually have empathy:. Nancy tries to save Oliver from the crime life and Charlie Bates was horrified by Nancy's death and even exposed Bill before changing his ways for good.” Speaker 2: ”Bill Sykes, on the other hand, was a villain I very much hated.” Speaker 1: ”**Agreed and I feel the same way**. They both represented redemption.. Fagin and Mr. Bumble were the villains you hate with a burning passion.. Fagin was a jerk who forced children to work as criminals without any care of how they feel, while Mr. Bumble was a fat hypocrite who saw the people of lower class as non-humans who deserved to be treated poorly.” Speaker 2: ”I didn't hate Fagin because even though he was a shameless crook who used kids to do his dirty work, he only partook in a victimless crime, abhorred violence, and didn't want to see the kids get hurt.” Speaker 1: ”And finally, Bill Sikes and Monks were the villains who were too epic to hate.” Speaker 2: ”Putting children in danger for your own gain without care for their well-being and threatening to murder them isn't epic in my book.”

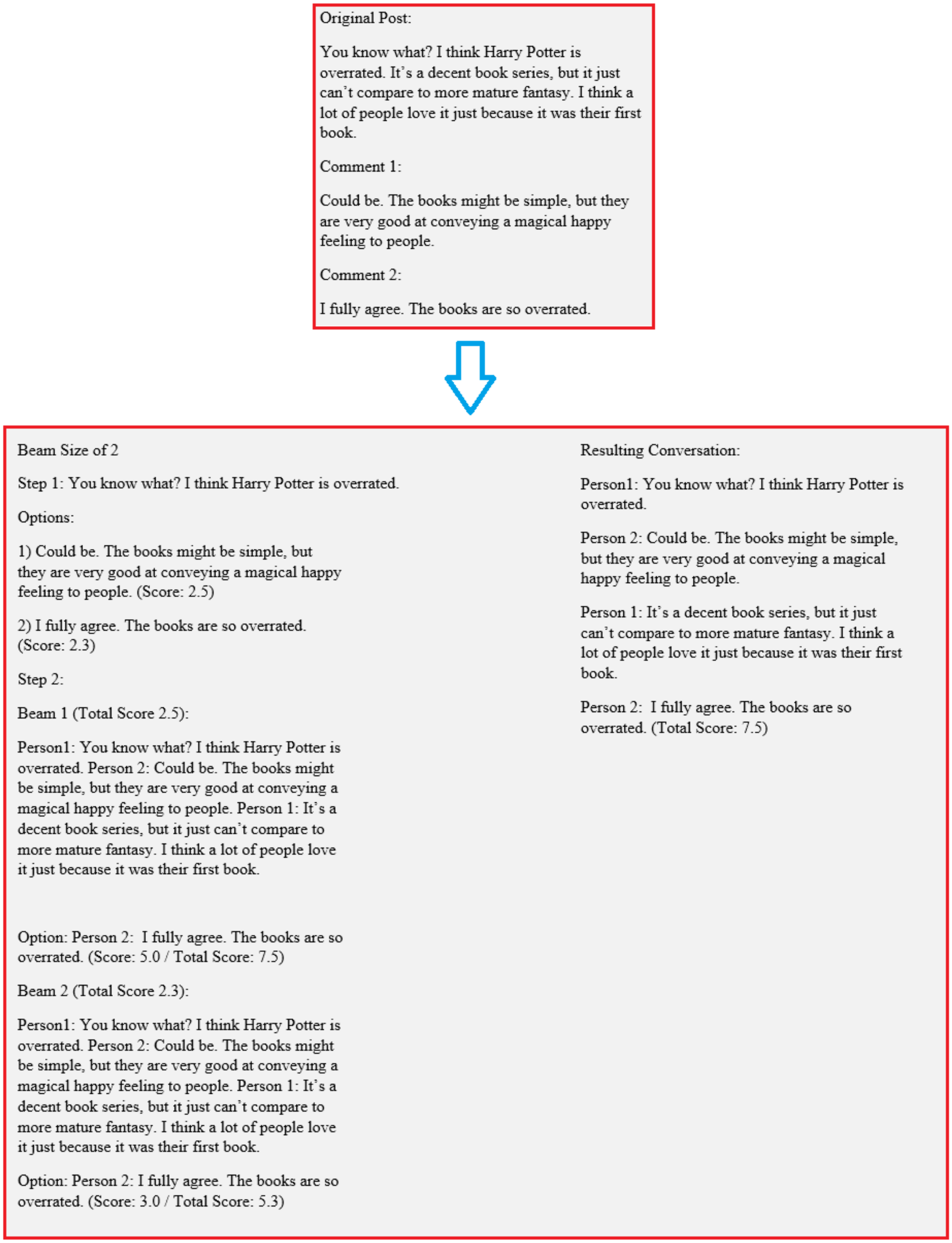
Figure 3.7: Example of how Threading improves conversations. Bold text is Threading addition

part of our work, we noticed we spent a lot of time developing dialogue that was built for specific popular topics - movies, books, sports, etc. One problem we found is that there is not a lot of data for some topics out there, with topic-focused datasets like Topical-Chat having only 8 topics. This is not a huge issue if we are training a chatbot that can talk about some topics, but for a truly natural experience, a chatbot should be able to talk about any topic. Actual humans not only have a huge knowledge base of the world they can utilize, but they can also intuit many things about new topics, allowing them to talk about them. For a model to both have a large knowledge base and be able to make intuitions about new topics, not only are more advanced neural approaches needed, but a much larger selection of dialogue data is required than is currently out there. We saw a need for not only larger amounts of data, but also an ability to on-demand create data about specific topics, enhancing deficient areas of chatbots. This paper is in part a response to the need we observed while working on Emora, our teams chatbot.

To give more context, during the Alexa Grand Prize we worked on several components of the chatbot (saliency propagation, efficiency improvements), but the most relevant one is us hand crafting dialogue for specific topics. This was a painstaking process of thinking through possible conversation paths, and creating logical structures to match to those situations, triggering appropriate responses. For many of the topics (like Olympics), it was simply not possible to train a neural model specializing in talking about those topics, as there was no high quality data existing for it. Our hope is that this project can alleviate that issue.

3.4 Collaboration

Over the course of this research, not only did we incorporate a large amount of coding-intensive approaches, we also had to manually evaluate a large amount of data (more on that in the next chapter). We talked with Dr. Choi when starting out on the project, and he suggested that it could be a joint-research project with Mack Hutsell, due to the the projects ambitious nature. That is what we did, and, while we both worked together on most of the approach, Mack focused more on some of the Reddit-side parts of the project, contributing more to comparing BERT vs DialogRPT [3], Comment handling, Beam Search, and Threading, while I focused on Post-filtering (GRADE), Pre-filtering, Reactions (seq2seq training / testing and BERT NSP reaction list), Blender and fine-tuning. For Core Flow (The main code of the final approach), we worked together. Where possible, this paper focuses less on Mack's work, but in many situations that is not practical, as it is a necessary part of the approach.



Original Post:

You know what? I think Harry Potter is overrated. It's a decent book series, but it just can't compare to more mature fantasy. I think a lot of people love it just because it was their first book.

Comment 1:

Could be. The books might be simple, but they are very good at conveying a magical happy feeling to people.

Comment 2:

I fully agree. The books are so overrated.

Beam Size of 2

Step 1: You know what? I think Harry Potter is overrated.

Options:

1) Could be. The books might be simple, but they are very good at conveying a magical happy feeling to people. (Score: 2.5)

2) I fully agree. The books are so overrated. (Score: 2.3)

Step 2:

Beam 1 (Total Score 2.5):

Person1: You know what? I think Harry Potter is overrated. Person 2: Could be. The books might be simple, but they are very good at conveying a magical happy feeling to people. Person 1: It's a decent book series, but it just can't compare to more mature fantasy. I think a lot of people love it just because it was their first book.

Option: Person 2: I fully agree. The books are so overrated. (Score: 5.0 / Total Score: 7.5)

Beam 2 (Total Score 2.3):

Person1: You know what? I think Harry Potter is overrated. Person 2: Could be. The books might be simple, but they are very good at conveying a magical happy feeling to people. Person 1: It's a decent book series, but it just can't compare to more mature fantasy. I think a lot of people love it just because it was their first book.

Option: Person 2: I fully agree. The books are so overrated. (Score: 3.0 / Total Score: 5.3)

Resulting Conversation:

Person1: You know what? I think Harry Potter is overrated.

Person 2: Could be. The books might be simple, but they are very good at conveying a magical happy feeling to people.

Person 1: It's a decent book series, but it just can't compare to more mature fantasy. I think a lot of people love it just because it was their first book.

Person 2: I fully agree. The books are so overrated. (Total Score: 7.5)

Figure 3.8: An example of beam search with our approach with beam of size n=2. On the left is the original post, on the right is the process and finished conversation. Note that the finished conversation would not have been possible if the best option was chosen at each step, and it scores higher than the greedy result.

Chapter 4

Experiments

4.1 Categorical Metric

Our initial categorical metric for manual evaluation was relatively simple. It rated a dialogue based on how much subjective ”sense” it made. See Table 4.1. We found that this metric did not accurately capture what we thought were significant differences in quality, as it failed to account for a human sense of how good interactions are. We moved on to the Continuous Metric for manual results, and only our initial tests were done with the categorical metric. Our main reasons for continuing to develop our metric was that 1) conversations we knew were more natural were not rated higher, due to them having a similar amount of explicit issues, and 2) the scale was off, as very few dialogues even in the highest quality datasets are perfect or only have a few insignificant issues, so the majority of conversations were below 4.

Table 4.1: Categorical Metric

Score	Conversation ”sense”
1	Nonsensical
2	Some parts make sense
3	Coherent but not human-like
4	Only a few small issues
5	Perfect natural dialogue

4.2 Continuous Metric

Table 4.2: Continuous Metric

Score	Conversation Quality
1	No good interactions
2	25% good interactions
3	50% good interactions
4	75% good interactions
5	100% good interactions

Our improved, continuous metric aimed to instead capture quantitatively how much of the conversation was good, and how much stood out as nonsensical. See Table 4.2. Not only did this metric more closely match our perceptions of quality, it also allowed us to more easily justify a certain score, by showing which interactions exactly we considered good and bad. As an added bonus, we could also far more easily give intermediate scores, with for instance 4.5 signifying 87.5% of interactions were good.

4.3 Manual Results

Manual Results (done by us) were used mainly for us to see which of our improvements had a positive effect, so the improved granularity of the continuous metric proved very useful here.

4.3.1 Categorical Metric

While we did not stick with the initial metric, it did provide some good information about how our approaches stacked up. As seen in Table 4.3, Beam search w/ the improvements detailed in our approach section did considerably better than the Base approach, while Blender offered no significant improvement (at the cost of being about 4x slower - see 4.4). From these preliminary results, we decided to focus on the beam search approach, working on more improvements for it, disregarding Blender. Blender was a significant part of our approach at this point, but the results (and our own analysis of the conversations) showed that it simply was not fit for this task.

Table 4.3: Categorical Metric Results

Approach	Score (100 conversations)
Base	1.65
Beam Search n=2 w/ improvements	2.57
Beam Search n=4 w/ improvements	2.53
Beam Search n=8 w/ improvements	2.61
Beam Search n=2 w/ Blender	2.61
Beam Search n=4 w/ Blender	2.61
Beam Search n=8 w/ Blender	2.63

Table 4.4: Categorical Metric Results Time Taken

Approach	Conversation Turns Per Second
Base	0.71
Beam Search n=2 w/ improvements	0.33
Beam Search n=4 w/ improvements	0.12
Beam Search n=8 w/ improvements	0.06
Beam Search n=2 w/ Blender	0.052
Beam Search n=4 w/ Blender	0.028
Beam Search n=8 w/ Blender	0.014

4.3.2 Continuous Metric Results

Note: for the continuous metric, we manually tested a lot more data, along with the metric being a bit more time-intensive to use, and therefore shortened the amount of posts we graded to 50. The results from this section can be seen in Table 4.5.

The continuous metric was our main way of looking at which of our approaches were more effective, as we felt it strongly correlated with human perceptions of conversation quality.

To begin with, our base approach showed a score of 2.95. This meant that roughly 50% of interactions looked like normal conversation interactions, and another 50% did not. Our base approach w/ improvements (Better NSP, Better Comments) showed a significant improvement at 3.26, demonstrating these improvements moved the model in a better direction.

We also looked at the effectiveness of Beam Search on our model, and found that with a beam of size n=2 we gained a significant improvement, while further beam growth only showed a small improvement at the cost of double the computational resources. From this we settled on a beam size of 2.

Final Beam Search is our Beam Search approach with all of our improvements, including looking at sequences of comments, along with us fixing some issues with Pre-Filtering that would occasionally interfere with the previous approach. It also showed a significant improvement over previous approaches, with a 4.12 score at beam size of n=2.

For our Fine-Tuned Model, we did not observe a significant improvement in score, however we decided that, since a Fine-Tuned model is an industry standard approach and the speed was exactly the same as non fine-tuned, we would use the model for our later Crowdworker testing.

Both reactions and threading showed a promising score increase to 4.24 and 4.26 respectively (especially considering how close we already to 5.0), and we decided to go with Threading, as not only did it score slightly higher, but we also felt that since it used comments from the post itself, it was more easily generalize-able to posts that our reaction list could not adapt to.

4.4 Crowdworker Results

For the final evaluation of our data, we decided to use Amazon Mechanical Turk. It is a platform where we hired human annotators to rate our data against other popular dialogue datasets. Because it is impartial humans evaluating our data, we wanted to

Table 4.5: Continuous Metric Results

Approach	Score (100 conversations)
Base	2.95
Base w/ improvements	3.26
Beam Search n=2 w/ improvements	3.42
Beam Search n=4 w/ improvements	3.5
Beam Search n=8 w/ improvements	3.45
Beam Search n=2 w/ improvements + GRADE	3.64
Final Beam Search n=1	3.76
Final Beam Search n=2	4.12
Final Beam Search n=2 + Fine Tuned Model	4.16
Final Beam Search n=2 + Fine Tuned Model + Reactions	4.24
Final Beam Search n=2 + Fine Tuned Model + Threading	4.26

use this to be the main benchmark of our results. Unfortunately, this turned out to be a mistake, so most of the data shown here is not that indicative of quality. We are still showing the results here (and the conclusions we had drawn) to illustrate the analysis we did on them.

We compared 3 popular dialogue datasets: Daily Dialogue, a high quality dataset of normal human conversations, Topical-Chat, high quality dialogue focused on specific topics, and MultiWOZ, a high quality dataset focused on annotating a large quantity of various human extremely specific conversations.

Because dialogue is far easier to compare than to rate on its own, we randomly paired one of our dialogues and matched it with a dialogue from one of these datasets, and asked the following question: is the first dialogue Significantly Less Natural, Less Natural, The Same, More Natural, or Significantly More Natural compared to the second? We had the task doubly annotated, which means every comparison was done twice by different annotators. For large scale data generation, we focused on books and movies topics, as those topics seemed to have the largest proportion of text based posts that we found (as opposed to links, images, etc.). We created 200 conversations from each. On average, we generated 0.05 conversation turns per second with our final approach.

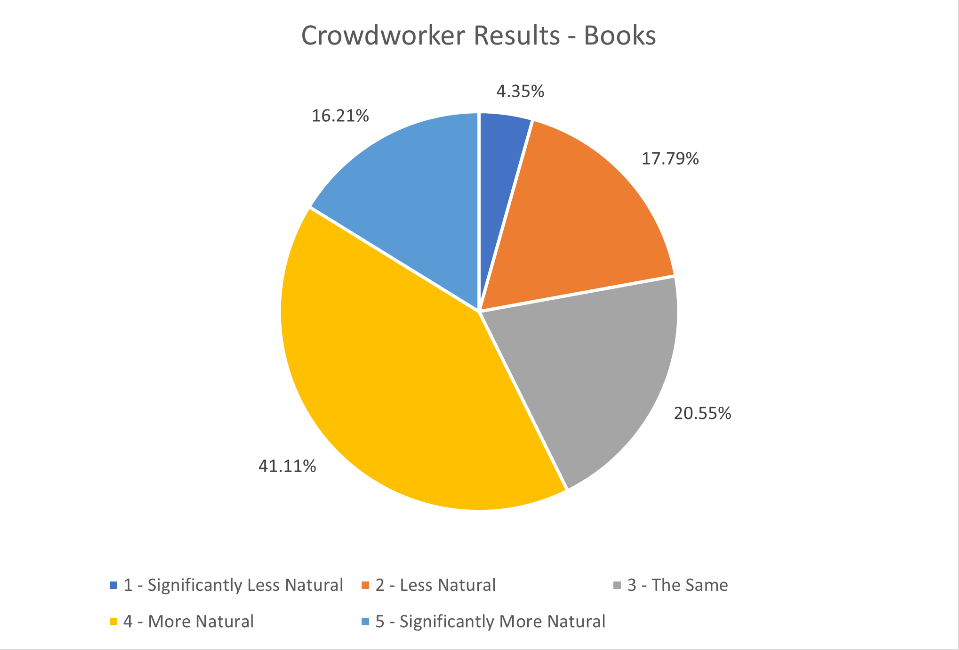


Figure 4.1: Crowdworker Results for Books dialogues against all datasets

Lastly, we only considered ratings that were given after a minute, as in our own testing it would take around 2 minutes to fairly read 2 dialogues and rate them. For the main data, see Figures 4.1 and 4.2. Our data looks to be significantly more natural for both our books and movies data, and by a surprisingly large amount.

Next, we looked at how our data compared against the other dialogue datasets on their own. We compared with 92 MultiWOZ conversations, and the results are in Figure 4.3, 174 conversations from DailyDialogue in Figure 4.4, and 134 conversations from Topical-Chat in Figure 4.5. What was surprising too us was that there is no clear better dataset, they all seemed to be roughly the same against ours - we had thought Daily Dialogue was of higher quality compared to the others.

Finally, we took a closer look at the Crowdworker data, and looked at how the two people annotating one dialogue matched up. Out of 400 dialogues, 172 dialogues have annotators who roughly agree (for instance that one is less natural, but maybe not on how much), and 122 dialogues have annotators who exactly match. If we take a look at coarse agreement (merge Significantly Less and Less, Significantly More and

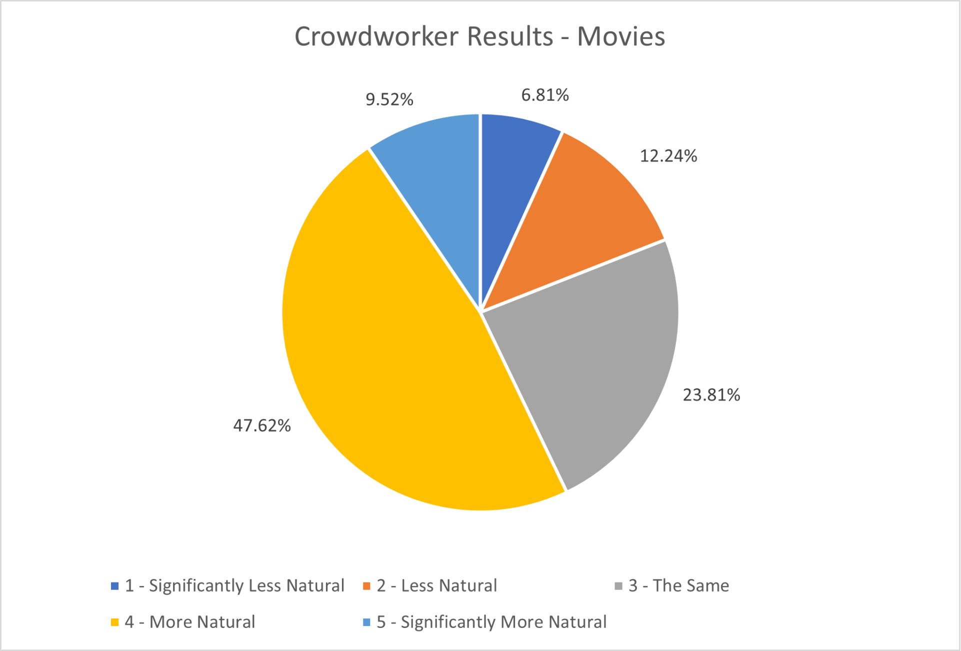


Figure 4.2: Crowdworker Results for Movies dialogues against all datasets

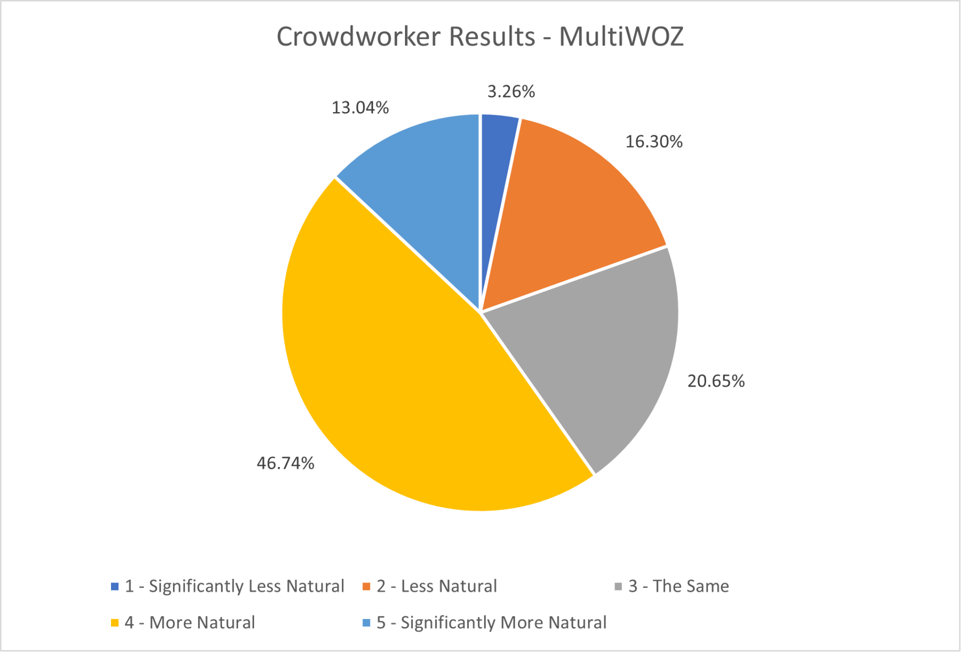


Figure 4.3: Crowdworker Results against all MultiWOZ

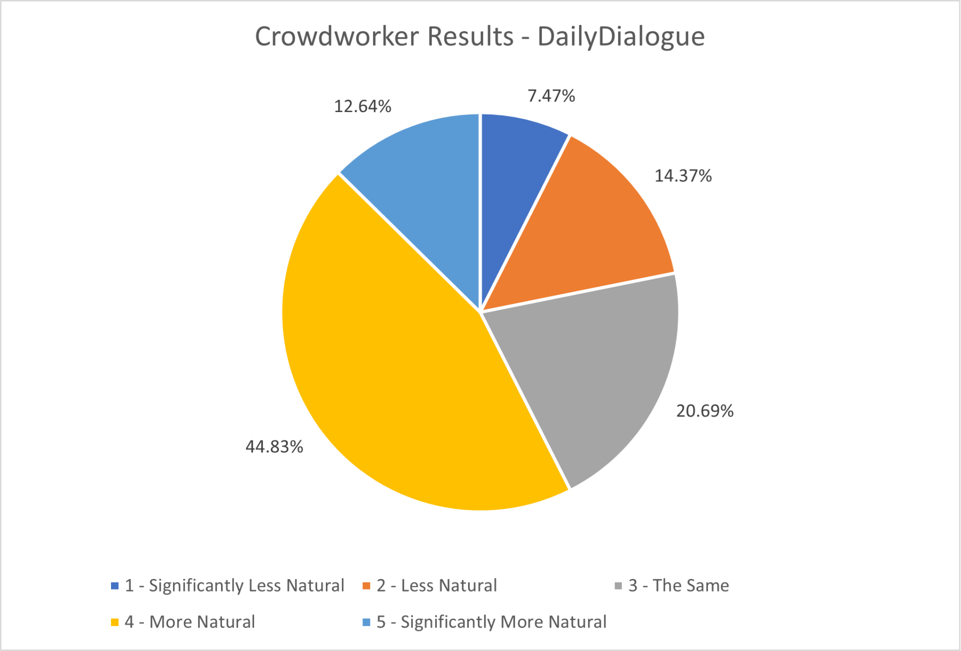


Figure 4.4: Crowdworker Results against all DailyDialogue

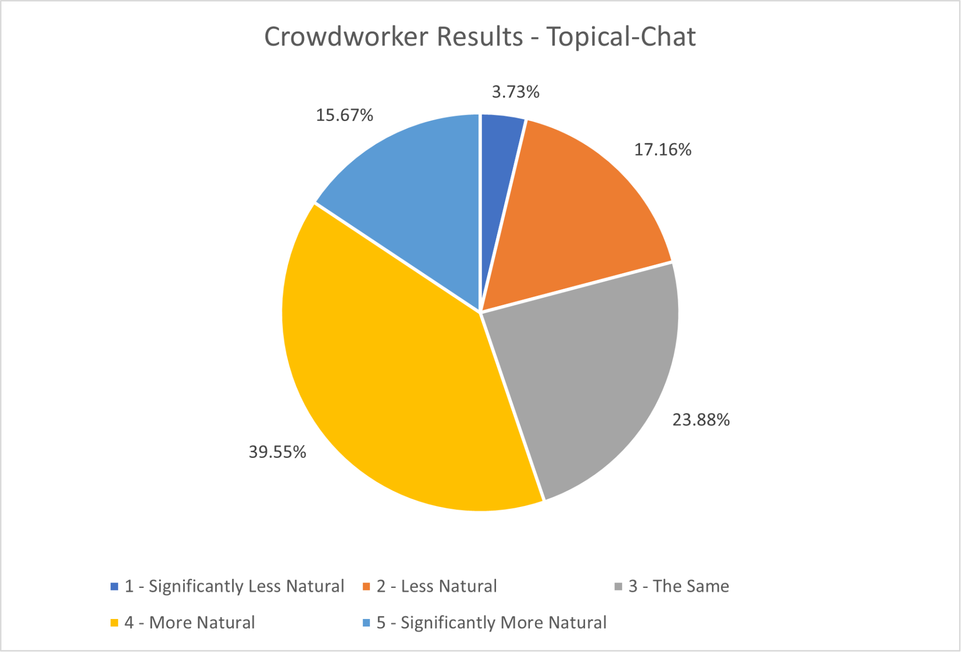


Figure 4.5: Crowdworker Results against Topical-Chat

More), our results are still roughly the same, as seen in table 4.6

Table 4.6: Crowdworker Results with Coarse Categories

Score	Amount
1 - Less Natural	9.88%
2 - The Same	12.21%
3 - More Natural	77.90%

Unfortunately, the Crowdworker results proved to be invalid. We doubly annotated our data in random order, where it we used a random number generator function in python to choose if our dialogue is seen first in the TURK task or second, i.e. if when someone selects "4 - More Natural", is ours considered more natural or less natural. During later experiments we ran a second TURK task to verify something we suspected: TURK users essentially always choose a higher number. In the first task, our conversations turned out to be placed on the left around 70% of the time due to the random method, and this led to the results we had. When we ran a subset of our task, but with the order flipped, our results were essentially the opposite. Crowdworkers proved to essentially randomly chose results, tending towards higher numbers, even with double annotation.

4.5 Final Expert Double Annotated Results

Of course, this was near to the end of our project, so we did not have the time to organize a large scale human annotation project, nor do we have the necessary resources to do so. What we decided to do was to personally double annotate 100 dialogues, with 50 from the Books dataset, and 50 from the movies dataset. While not a perfect amount, we believed it was enough to show at least some measure of quality, considering we had 2 NLP undergraduate students manually double annotate 100 dialogues. The results are in Figure 4.6 and Table 4.7. While the results are not as drastic as with the Crowdworker data, it still shows that our data is significantly

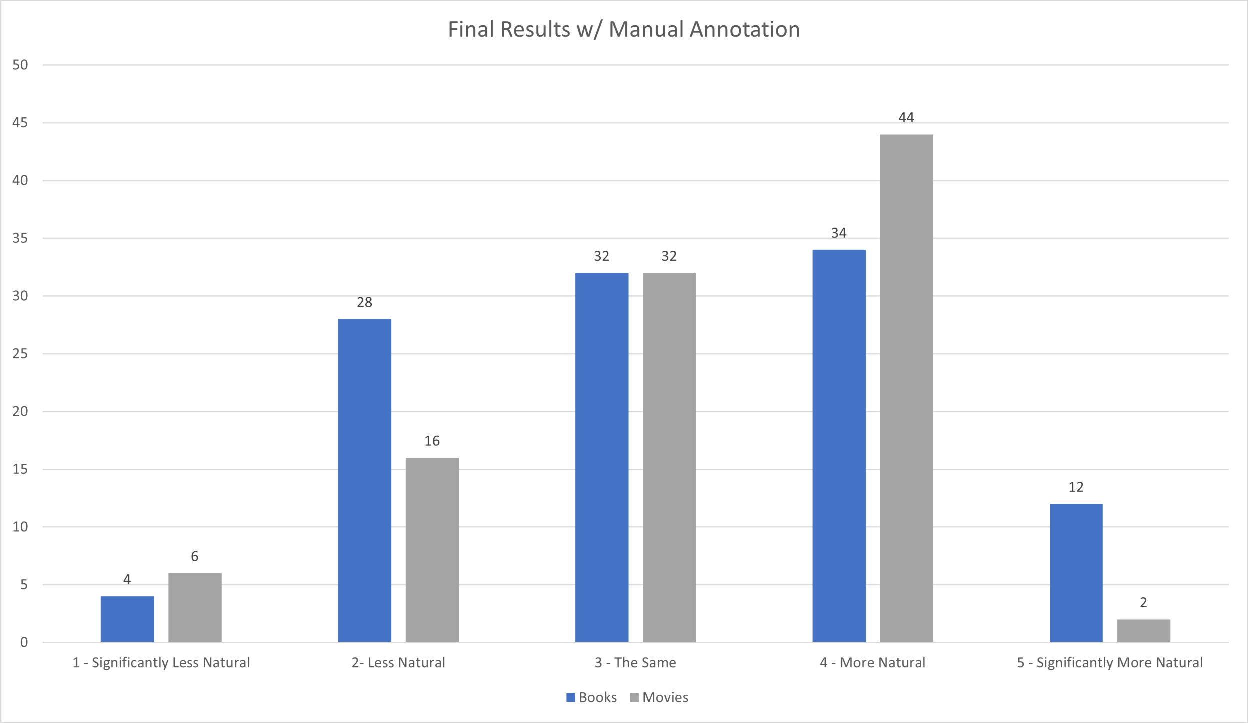


Figure 4.6: The final results from expert annotation.

more natural. 27% of our data is ”as natural” as the datasets combined, and 46% ”more natural”, with only 27% ”less natural”, a clear and significant advantage, with the movies topic having a slightly favourable amount of ”the same” instead of ”less natural” compared to the books topic.

Table 4.7: Coarse Final Results (More Natural means ours is more natural)

Score	Books	Movies
1 - Less Natural	32%	22%
2 - The Same	22%	32%
3 - More Natural	46%	46%

Finally, the comparisons against specific datasets are in Tables 4.8, 4.9, and 4.10. Of course, we had a much smaller amount of conversations per dataset due to expert annotation, but if we attempt to draw limited conclusions from this, our data is roughly equal in quality to Daily Dialogue and MultiWOZ, and significantly better than Topical-Chat.

Table 4.8: Coarse Final Results against Daily Dialogue

Score	Percentage
1 - Less Natural	35.14%
2 - The Same	27.02%
3 - More Natural	37.84%

Table 4.9: Coarse Final Results against MultiWOZ

Score	Percentage
1 - Less Natural	23.08%
2 - The Same	42.30%
3 - More Natural	34.62%

Table 4.10: Coarse Final Results against Topical-Chat

Score	Percentage
1 - Less Natural	14.81%
2 - The Same	14.81%
3 - More Natural	70.38%

4.6 Analysis

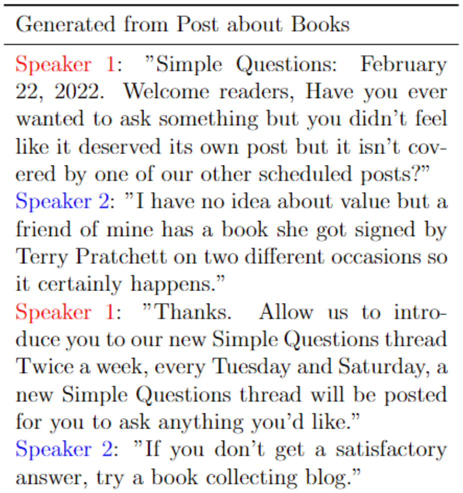
For analysis, we decided to look at some of our lower rated conversations to determine the weaknesses of our approach. Something we noticed relatively often was, as seen in Figure 4.7, the original post itself is often low quality. The only thing we can really do here is to more comprehensively filter for grammatical correctness / unnatural language, something we intend to do in the future.

Another type of common weakness can be seen in Figure 4.8. The very first words of Speaker 1 are ”Simple Questions: February 22, 2022. Welcome readers”. This is part of a special post made to encourage discussion in the comments on ”r/books”. While it is relatively easy to filter out all ”Simple Questions” posts (or only use the comments, not the post itself), there are many types of these posts of Reddit. This is a significant problem, and our current plan is to train a model that can predict if this is the case based on the text and punctuation of the post.

Finally, we noticed posts like Figure 4.9. At the very end, Speaker 2 seems to agree with Speaker 1 twice, with the second time appearing to be first time he is convinced. He first says ”Actually, I second this”, and then ”You know what, I actually think you’re right”. We looked into this, and these types of issues arise from there being multiple commenters, both at some point changing their minds and agreeing with the main post. Our model thinks at both points that is a valid thing to do, and appears to struggle with remembering that Speaker 2 already changed his mind. This is a very hard problem to solve, and in future work our plans are to try and prevent duplicate statements like this by trying to model agreement/disagreement better, but this would be very far out for us in terms of priorities. Other notes we took from this analysis was that we should stay away from politically charged subreddits, as those almost never translated into good dialogues, in large part due to their confrontational nature. We

Generated from Post about Financial Aid
Speaker 1: ”Underwhelming book scare me and I read too many reviews, before choosing a book! I don’t know exactly why... but I fear often about books which are underwhelming, the content, conclusion, characters etc.” Speaker 2: ”I do try kindle sample..that is a good idea to continue.” Speaker 1: ”Whenever I am looking to choose a new book to read, I tend to read reviews from multiple sites(popular newspaper articles or Redditt) to make sure I don’t get disappointed reading that book and feel sad that I wasted precious time.” Speaker 2: ”I wander around book stores and just look at titles and covers, read the blurb, the first page and some random pages in the book. I rarely read reviews as I find they depend too much on the reviewer’s taste, age, interests etc.” Speaker 1: ”Haha. This most often, ruins the pleasure of reading a book, as I would know the book outline ahead and I feel like I am ruining my book experience.” Speaker 2: ”I do this even with authors that I love because there is the odd book or two of theirs that would do nothing for me. Reviews are very subjective. I’ve stopped writing them for books I’ve read.”

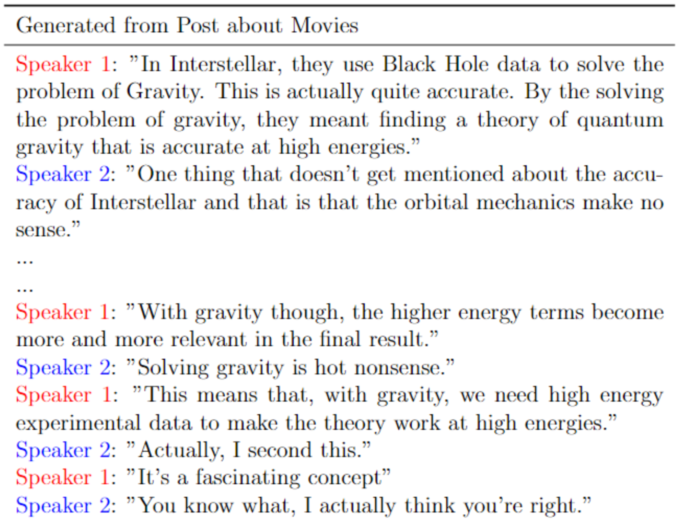
Figure 4.7: Low Rated Dialogue



Generated from Post about Books

Speaker 1: ”Simple Questions: February 22, 2022. Welcome readers, Have you ever wanted to ask something but you didn’t feel like it deserved its own post but it isn’t covered by one of our other scheduled posts?”
Speaker 2: ”I have no idea about value but a friend of mine has a book she got signed by Terry Pratchett on two different occasions so it certainly happens.”
Speaker 1: ”Thanks. Allow us to introduce you to our new Simple Questions thread Twice a week, every Tuesday and Saturday, a new Simple Questions thread will be posted for you to ask anything you’d like.”
Speaker 2: ”If you don’t get a satisfactory answer, try a book collecting blog.”

Figure 4.8: Low Rated Dialogue



Generated from Post about Movies

Speaker 1: ”In Interstellar, they use Black Hole data to solve the problem of Gravity. This is actually quite accurate. By the solving the problem of gravity, they meant finding a theory of quantum gravity that is accurate at high energies.”
Speaker 2: ”One thing that doesn’t get mentioned about the accuracy of Interstellar and that is that the orbital mechanics make no sense.”
...
...
Speaker 1: ”With gravity though, the higher energy terms become more and more relevant in the final result.”
Speaker 2: ”Solving gravity is hot nonsense.”
Speaker 1: ”This means that, with gravity, we need high energy experimental data to make the theory work at high energies.”
Speaker 2: ”Actually, I second this.”
Speaker 1: ”It’s a fascinating concept”
Speaker 2: ”You know what, I actually think you’re right.”

Figure 4.9: Low Rated Dialogue (abbreviated for space)

also noted that some subreddits seemed to have, on average, better material for us to work with. On ”r/books”, users typically had more natural and eloquent language, while, for example, ”r/funny” did not.

Chapter 5

Conclusion

So, overall, we can see that our approach appears to be giving data on par or superior to popular high-quality dialogue datasets - our dialogues are rated ”more natural” 46% of the time, and ”more natural” or ”the same” 73% of the time. Our results are backed up by our own internal ratings from earlier in the project, and while we were unable to use impartial Crowdworker data, our more robust doubly annotated expert evaluation backs this up. The automatic nature of our approach means that, not only is our data of relatively high quality, it is also effectively unlimited, or at least only limited by the scope of Reddit (300 million Reddit Posts are created every year). A specific interesting comparison is against Topical-Chat. Just like our dataset, it is focused on having people talk about specific topics, and, given our data can do the same thanks to our using data from specific subreddits, the fact that our dataset rates very highly against it (70% ”More Natural” and 84% ”The Same” or ”More Natural”) means our model could be used to create a higher quality topic focused dialogue dataset. Our approach allows us to create a dialogue dataset that is highly focused on any given topic, making it extremely useful for those who need to train for a very specific task. We have shown that we effectively utilize the nature of Reddit data to create natural, multi-turn dialogue that is at the very least on par with human created dialogue.

5.1 Future Work

Naturally, we would like to conduct impartial evaluation of our dataset, something we thought TURK would allow us to do. We also intend to create more dialogues, and compare with all of the most popular dialogue datasets, like Switchboard. Getting more data points across more datasets will allow us to further demonstrate the strengths of our model, as well as illustrate what its weaknesses are and what we should work on. Apart from that, one of the things we want to look at is using other online data repositories, like Twitter or Facebook. Perhaps our data might be better suited for some topics depending on the website, for example Facebook data might be able to more accurately represent a conversation between friends instead of strangers. While we evaluated several forums at the beginning of our project, there are always more options to consider. On the other hand, Twitter seems to have fewer deep conversation threads (in our experience).

In the immediate future, we plan to make a few improvements to our pre-filtering process, to improve the quality of language our model has to work with. We also want to see whether we can neurally augment more of the filtration process, as at the moment only post-filtering is neural (GRADE). We also plan to find more applications that require topic specific datasets, and comparing our data to the data already available in that field (if any meaningful amount even exists).

Appendix A

Pseudocode (Blender Approach)

```
Helper Functions

getScore(prompt, option)
    tokenize prompt and option variables
    Get BERT NSP score for prompt, option
    Return positive class prediction score minus negative class prediction score

getMostLikely(curr_state, options)
    For each option in options
     call getScore
     If this is the highest score yet, then store the score and the option
    Return highest score and the corresponding option

getBlenderResponse(curr_state, process)
    Read in from stdout of process variable into a line variable
    Wait until \[BlenderBot2Fid" is not in the line
    Wait until \Enter Your Message:" is in the line
    Write the last sentence from curr_state variable to stdin of process variable
    Write \[DONE]" to stdin of process variable
    Read in from stdout of process variable
    Wait until \... preparing new chat..." is not in the line
    Return line

blenderInitialize(process)
    Read in from process variable stdout into line variable
    Wait until \creating task(s): interactive" in line variable
    Write \[DONE]" into process variable stdin
    Read in from process variable stdout into line variable
    Wait until \... preparing new chat..." is in the line variable
    Return

getComments(post)
    Initialize comment list
    For comment in post:
     Get combinations of comment up to size 3
     Add them all to the comments list
    Return comments


getRandomPost
    Get a list of subreddit directory names
    Generate a random integer between 0 and the number of subreddits
    Get the corresponding subreddit directory name
    Get a list of the post names in that subreddit
    Generate a random integer between 0 and the number of posts
    Load the corresponding post
    Return the subreddit name and the post


main()
    Start a subprocess with blender and save it to a process variable
    Call blenderInitialize(process)
    Call getRandomPost() and store the data into a post_data variable
    If the length of the post is over 1000 characters or it has more than ten sent
    Split the post's text into sentences
    Call getComments(post_data)
    Initialize a curr_state list
    For curr_sentence in split_post
     Calculate score of BERT NSP from previous curr_state to curr_state + curr_sen
     IF the quality is not high enough (<12), then if using Blender, prepend a Ble
    Append curr_sentence to curr_state
    Call getMostLikely(curr_state, comments + the next curr_sentence) and store re
    While the returned value of getMostLikely (next[1]) is the next curr_sentence
     Add the returned value to the current entry of curr_state (don't add a new el
     Call getMostLikely again and store the result in the next variable
    Append next[1] to the curr_state list
```

Appendix B

Example Conversations

B.1 Example BERT NSP Baseline Post

Speaker 1) ”I mean in terms of a lack of energy and just so fatigued all the time”,

Speaker 2) ”I feel like I haven’t retained anything”,

Speaker 1) ”I manage to get a good amount of sleep every night, yet I’m still tired throughout the day”,

Speaker 2) ”I work and only go to school part-time and that already tires me out”,

Speaker 1) ”It’s been harder for me to do my assignments without me feeling drowsy and just wanted to crawl back into bed and do nothing. I have an exam tomorrow morning at 8am and I just don’t have the energy to study for it even though I wrote notes on the chapters that will be on the test and I also studied previously”,

Speaker 2) ”I’m super ADHD and anxious so it’s hard to focus”,

Speaker 1) ”But I just feel so burnt out and tired that I don’t want to do any more schoolwork”,

Speaker 2) ”Just doing the work/exams and then not retaining any of the information”, Speaker 1) ”I know for a fact I’m not the only one experiencing this, but I figured I’d just throw this on Reddit and see what others have to say”,

Speaker 2) ”I also cant focus in class without getting antsy so i put on music or reddit videos”

B.2 Example DialogRPT Baseline Post

:

Speaker 1)”I’m sitting on my porch at 2am with a bunch of my colleagues”,

Speaker 2)”I’m a tenured STEM professor with a PhD and I make $61k/year”,

Speaker 1)”We are all professors and we are all drunk and we all love our students so much and want you to know that we are real human beings that empathize with everything you’re going through”,

Speaker 2)”Nobody becomes a professor for the money”

B.3 Example BERT NSP Beam 2 Post

:

Speaker 1) ”College readings are impossible. Am I the only one who get super frustrated trying to read an article for a college assingment where the author focuses more on pumping out as many five-dollar words in a row as they can than actual readability or comprehension? (I know, ironic given my run-on, but cut me some slack”,

Speaker 2) ”Yes, I have at least 3, 7-10 page readings I have to write about a week”,

Speaker 1) ”I’ve been reading anthropology articles all day)”,

Speaker 2) ”Well this is what happens when you focus on STEM without the Humanities, you end up with scientists that are unable to communicate their knowledge to the vast majority of people”,

Speaker 1) ”Anyways, if I have to google every other word, I’m probably not going to finish the article”,

Speaker 2) ”What I do, is I just read the entire thing without stopping, maybe taking occasional notes on what I’m getting, and then see if I come come up with a general summary”,

Speaker 1) ”I get it, it’s nice to use complicated words that express your meaning better than simpler ones, I like to use them too”,

Speaker 2) ”Also, looking up whatever you’re reading + summary (although a lot of the stuff I read, and perhaps you read, is so down the rabbit hole it won’t be on the internet) and then when you think back to the details of the paper it will all make sense hopefully”,

Speaker 1) ”Still, as a college student I would at least like to be able to read and comprehend what I’m supposed too in a reasonable amount of time”,

Speaker 2) “They’re not impossible if it’s a PDF and you’re not having a quiz on it only discussions post but at the end of the day all you really need to do is skim it and read the conclusion and abstract and if you don’t have a quiz Central f the question that’s what I always do”,

Speaker 1) ”It’s really annoying and I wish college professors would choose articles with simplified wording for complex topics”,

Speaker 2) ”GOD YESlike i understand it’s a formal paper or whatever and even if i know what the words mean, there’s nothing that fried my brain more than an overly wordy sentence that uses giant words just to sound smarter”

B.4 Example BERT NSP Beam 4 Post

:

Speaker 1) ”College readings are impossible. Am I the only one who get super frustrated trying to read an article for a college assingment where the author focuses more on pumping out as many five-dollar words in a row as they can than actual readability or comprehension? (I know, ironic given my run-on, but cut me some slack”,

Speaker 2) ”Yes, I have at least 3, 7-10 page readings I have to write about a week”,

Speaker 1) ”I've been reading anthropology articles all day)”,

Speaker 2) ”Well this is what happens when you focus on STEM without the Humanities, you end up with scientists that are unable to communicate their knowledge to the vast majority of people”,

Speaker 1) ”Anyways, if I have to google every other word, I'm probably not going to finish the article”,

Speaker 2) ”What I do, is I just read the entire thing without stopping, maybe taking occasional notes on what I'm getting, and then see if I come come up with a general summary”,

Speaker 1) ”I get it, it's nice to use complicated words that express your meaning better than simpler ones, I like to use them too”,

Speaker 2) ”Also, looking up whatever you're reading + summary (although a lot of the stuff I read, and perhaps you read, is so down the rabbit hole it won't be on the internet) and then when you think back to the details of the paper it will all make sense hopefully”,

Speaker 1) ”Still, as a college student I would at least like to be able to read and comprehend what I'm supposed too in a reasonable amount of time”,

Speaker 2) ”GOD YESlike i understand it's a formal paper or whatever and even if i know what the words mean, there's nothing that fried my brain more than an overly wordy sentence that uses giant words just to sound smarter”, Speaker 1) ”It’s really annoying and I wish college professors would choose articles with simplified wording for complex topics”,

Speaker 2) ”They’re not impossible if it’s a PDF and you’re not having a quiz on it only discussions post but at the end of the day all you really need to do is skim it and read the conclusion and abstract and if you don’t have a quiz Central f the question that’s what I always do”

B.5 Example BERT NSP Beam 8 Post

:

Speaker 1) ”College readings are impossible. Am I the only one who get super frustrated trying to read an article for a college assingment where the author focuses more on pumping out as many five-dollar words in a row as they can than actual readability or comprehension? (I know, ironic given my run-on, but cut me some slack”,

Speaker 2) ”Yes, I have at least 3, 7-10 page readings I have to write about a week”,

Speaker 1) ”I’ve been reading anthropology articles all day)”,

Speaker 2) ”Well this is what happens when you focus on STEM without the Humanities, you end up with scientists that are unable to communicate their knowledge to the vast majority of people”,

Speaker 1) ”Anyways, if I have to google every other word, I’m probably not going to finish the article”,

Speaker 2) ”What I do, is I just read the entire thing without stopping, maybe taking occasional notes on what I’m getting, and then see if I come come up with a general summary”,

Speaker 1) ”I get it, it’s nice to use complicated words that express your meaning better than simpler ones, I like to use them too",

Speaker 2) "Also, looking up whatever you're reading + summary (although a lot of the stuff I read, and perhaps you read, is so down the rabbit hole it won't be on the internet) and then when you think back to the details of the paper it will all make sense hopefully",

Speaker 1) "Still, as a college student I would at least like to be able to read and comprehend what I'm supposed too in a reasonable amount of time",

Speaker 2) "GOD YESlike i understand it's a formal paper or whatever and even if i know what the words mean, there's nothing that fried my brain more than an overly wordy sentence that uses giant words just to sound smarter",

Speaker 1) "It's really annoying and I wish college professors would choose articles with simplified wording for complex topics",

Speaker 2) "They're not impossible if it's a PDF and you're not having a quiz on it only discussions post but at the end of the day all you really need to do is skim it and read the conclusion and abstract and if you don't have a quiz Central f the question that's what I always do"

B.6 Example BERT NSP + Blender Beam 2 Post

:

Speaker 1) "Existential dread of starting a PhD in the current world climate. Saw a post a little while ago expressing a sort of existential dread about entering grad school with the current state/climate of the world and I share that worry",

Speaker 2) "Saw a post a little while ago expressing a sort of existential dread about entering grad school with the current state/climate of the world and I share that worry",

Speaker 1) "I'm normally an optimist about this kind of stuff, but the state of things (climate change, etc",

Speaker 2) "The last post about this anxiety really laid out some solid evidence that the likelihood of a catastrophic global climate apocalypse is extremely unlikely in our lifetime",

Speaker 1) ") continues to get worse and I'm really getting concerned whether spending 5-10 years on a PhD is worth it at this point",

Speaker 2) "That is, if you definitely would want to do a PhD if climate change didn't exist, just go for it! Who knows what future trajectory we'll all end up on",

Speaker 1) "I definitely have a lean towards industry as of now, if that has an influence in this",

Speaker 2) "This is why I'm thinking of doing grad school research on an environmental or climate change related topic as a way of doing what I love and doing something to help our planet",

Speaker 1) "I'm entering grad school with the intention of getting a PhD (epidemiology/biostats), but now I'm wondering if I should just get my Master's and move on with my life/career considering there's a very real possibility we could be in chaos in a few years",

Speaker 2) "I think to myself in regards to climate change, what would I be doing if I weren't in grad school",

Speaker 1) "I love my research to death and would love to see it through",

Speaker 2) "Would working a few extra years really matter in the end to protect myself from climate change? If food supply lines get disrupted (which they will to some degree at some point), what do I do then? The way I see it, I am betting on scientists, activists, and humanity prevailing (not an easy bet) which means continuing with grad school and getting involved in activism outreach as much as I can"

B.7 Example BERT NSP + Blender Beam 4 Post

:

Speaker 1) ”Existential dread of starting a PhD in the current world climate. Saw a post a little while ago expressing a sort of existential dread about entering grad school with the current state/climate of the world and I share that worry”,

Speaker 2) ”Saw a post a little while ago expressing a sort of existential dread about entering grad school with the current state/climate of the world and I share that worry”,

Speaker 1) ”I’m normally an optimist about this kind of stuff, but the state of things (climate change, etc”,

Speaker 2) ”The last post about this anxiety really laid out some solid evidence that the likelihood of a catastrophic global climate apocalypse is extremely unlikely in our lifetime”,

Speaker 1) ”) continues to get worse and I’m really getting concerned whether spending 5-10 years on a PhD is worth it at this point”,

Speaker 2) ”That is, if you definitely would want to do a PhD if climate change didn’t exist, just go for it! Who knows what future trajectory we’ll all end up on”,

Speaker 1) ”I definitely have a lean towards industry as of now, if that has an influence in this”,

Speaker 2) ”This is why I’m thinking of doing grad school research on an environmental or climate change related topic as a way of doing what I love and doing something to help our planet”,

Speaker 1) ”I’m entering grad school with the intention of getting a PhD (epidemiology/biostats), but now I’m wondering if I should just get my Master’s and move on with my life/career considering there’s a very real possibility we could be in chaos in a few years”,

Speaker 2) ”I think to myself in regards to climate change, what would I be doing if I weren't in grad school",

Speaker 1) "I love my research to death and would love to see it through",

Speaker 2) "Would working a few extra years really matter in the end to protect myself from climate change? If food supply lines get disrupted (which they will to some degree at some point), what do I do then? The way I see it, I am betting on scientists, activists, and humanity prevailing (not an easy bet) which means continuing with grad school and getting involved in activism outreach as much as I can"

B.8 Example BERT NSP + Blender Beam 8 Post

:

Speaker 1) "Existential dread of starting a PhD in the current world climate. Saw a post a little while ago expressing a sort of existential dread about entering grad school with the current state/climate of the world and I share that worry",

Speaker 2) "Saw a post a little while ago expressing a sort of existential dread about entering grad school with the current state/climate of the world and I share that worry",

Speaker 1) "I'm normally an optimist about this kind of stuff, but the state of things (climate change, etc",

Speaker 2) "The last post about this anxiety really laid out some solid evidence that the likelihood of a catastrophic global climate apocalypse is extremely unlikely in our lifetime",

Speaker 1) ") continues to get worse and I'm really getting concerned whether spending 5-10 years on a PhD is worth it at this point",

Speaker 2) "That is, if you definitely would want to do a PhD if climate change didn't exist, just go for it! Who knows what future trajectory we'll all end up on",

Speaker 1) "I definitely have a lean towards industry as of now, if that has an influence in this”,

Speaker 2) ”This is why I’m thinking of doing grad school research on an environmental or climate change related topic as a way of doing what I love and doing something to help our planet”,

Speaker 1) ”I’m entering grad school with the intention of getting a PhD (epidemiology/biostats), but now I’m wondering if I should just get my Master’s and move on with my life/career considering there’s a very real possibility we could be in chaos in a few years”,

Speaker 2) ”I think to myself in regards to climate change, what would I be doing if I weren’t in grad school”,

Speaker 1) ”I love my research to death and would love to see it through”,

Speaker 2) ”Would working a few extra years really matter in the end to protect myself from climate change? If food supply lines get disrupted (which they will to some degree at some point), what do I do then? The way I see it, I am betting on scientists, activists, and humanity prevailing (not an easy bet) which means continuing with grad school and getting involved in activism outreach as much as I can”

Bibliography

[1] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *CoRR*, abs/1810.00278, 2018. URL `http://arxiv.org/abs/1810.00278`.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[3] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. Dialogue response ranking training with large-scale human feedback data. *CoRR*, abs/2009.06978, 2020. URL `https://arxiv.org/abs/2009.06978`.

[4] J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1, 1992. doi: 10.1109/ICASSP.1992.225858.

[5] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079.

[6] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. *CoRR*, abs/2010.03994, 2020. URL https://arxiv.org/abs/2010.03994.

[7] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *CoRR*, abs/1710.03957, 2017. URL http://arxiv.org/abs/1710.03957.

[8] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. *CoRR*, abs/2107.07567, 2021. URL https://arxiv.org/abs/2107.07567.

[9] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[10] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536, 2019. URL http://arxiv.org/abs/1911.00536.