Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an undergraduate degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

William Mack Hutsell	04/12/2022
Signature	Date

Automatic Generation of Multi-turn Dialogues from Reddit

By

William Mack Hutsell

Jinho D. Choi, Ph.D.
Advisor

Computer Science

Jinho D. Choi, Ph.D.
Advisor

Lauren Klein, Ph.D.
Committee Member

Ting Li, Ph.D.
Committee Member

2022

Automatic Generation of Multi-turn Dialogues from Reddit

By

William Mack Hutsell
A.A., Oxford College, GA, 2020

Advisor: Jinho D. Choi, Ph.D.

An abstract of
A thesis submitted to the Faculty of Emory College of Arts and Science
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors
Computer Science
2022

Abstract

Automatic Generation of Multi-turn Dialogues from Reddit
By William Mack Hutsell

High-quality multi-turn dialogue datasets are a scarce commodity in the field of Natural Language Processing, and with the recent rise of chat bots powered by seq2seq models that train on these datasets, they have become more important than ever. This thesis describes work done on a model built to deconstruct Reddit posts and sequence the fragments to create high-quality, multi-turn, topic-specific conversations. The model works by using a post's content as a beginning framework for a single speaker's statements in a conversation, filling in the second speaker's utterances with comments left on the same post. A dialogue dataset with 951 dialogues was generated using this method comprising conversations across two topics: movies and books. This dataset, HuHu, was then manually evaluated against DailyDialog, Topical-Chat, and MultiWOZ, three good-quality datasets with ∼10,000 dialogues constructed in varying ways. The results showed that our generated dialogues were overall considered more natural in 46% of cases and considered at least as natural in 73% of comparisons. This is an incredible result given that our model can generate millions of dialogues across any number of topics, limited only by the number of related Reddit posts. Future work in the task of dialogue assembly models appears to be very promising and could result in dialogues at a near-human level within the near future.

Automatic Generation of Multi-turn Dialogues from Reddit

By

William Mack Hutsell
A.A., Oxford College, GA, 2020

Advisor: Jinho D. Choi, Ph.D.

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

First of all, I'd like to thank my advisor Dr. Jinho D. Choi. Coming from the Oxford College campus and transitioning to the Atlanta campus during COVID, I never expected to be able to pursue an honors thesis. Joining the Emory Natural Language Processing Lab under Dr. Choi's direction cultivated my passion for computer science and language, and I gained irreplaceable experience during my time there. Additionally, Dr. Choi's guidance throughout the development of my thesis was invaluable.

I'd also like to thank my committee members, Dr. Ting Li and Dr. Lauren Klein. Dr. Li was my first computer science professor at Emory. She was the reason I decided to stick with the major, and her instruction was indispensable. Dr. Klein is someone I have worked with over the last year on several projects. Each has greatly expanded my understanding of what it means to responsibly create and apply technology.

Finally, I'd like to give thanks to my friends and family; their support has made writing this thesis possible.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

So I find words I never thought to speak
For last year's words belong to last year's language.
See, they return, and bring us with them.

T.S. Eliot, *Little Gidding*

1.1 The Importance of Dialogue

Dialogue is the chief mediator of daily human life. It is one of the most basic yet complex interactions humans can partake in — a rich interplay of language, differing knowledge bases, and intents, a two-agent attempt at the creation of a single understanding. Because of its prominence in daily life, dialogue is a sought-after capability for models within the Natural Language Processing (NLP) field.

The applications for models capable of human-like dialogue are endless. Machines could explain the significance of data they had measured; online bots could help people find targeted information on the web; machines could help others practice new languages. Beyond basic applications such as these, a chat bot could be specially designed to speak with those living with depression or perhaps provide targeted reminders for anyone living with memory-related diseases. The goal, of course, is not

for chat bots to replace humans; these sorts of applications could serve as a safety net when actual human interaction is not present.

However, dialogue is an extraordinarily difficult phenomenon to accurately model. To compose syntactically correct, semantically seamless language within a dialogue requires managing real-world knowledge, intelligently processing the context of the dialogue, and producing natural language. Within the world of computers, each of these tasks, trivial for most humans, is nearly unapproachable in difficulty to complete at a human level.

1.2 The Dialogue Assembly Task

Many modern chat bots train on dialogue data, collections of conversations of various lengths and topics; but as dialogue is difficult to produce, there is a scarcity of high-quantity and high-quality dialogue datasets. From this scarcity comes the necessity of a model that is able to assemble these sorts of dialogues from resources that are available in large quantities. This can be thought of as a dialogue-assembly task. For this thesis, this task will be addressed via assembling multi-turn dialogues from Reddit, an online bulletin-board social media.

A key component of this task is the ability to provide high-quality, multi-turn, topic-specific dialogue data in high quantities. However, it is also important that the model that addresses this task is extensible and easily modifiable. While creating or sourcing one dialogue dataset that fits our parameters is good, what is clearly better is the ability to at-will generate such datasets or slight variants of them using a model. The use of Reddit posts provides varied and endless content, leaving us with only the need to design a flexible model that can create high-quality and multi-turn dialogues out of the posts.

1.3 Thesis Statement

The work of this thesis has a dual purpose. By providing a mechanism by which high-quality and topic-specific conversation data can be automatically generated, we expect to increase the quantity of high-impact dialogue data available for the use of training seq2seq or other dialogue models; this data, of course, could be used for any NLP applications, not necessarily just training dialogue models. We also expect to provide a flexible mechanism by which researchers can generate topic-specific conversations for their own use. Additionally, by exploring and obtaining good results within the relatively new sub-task of automatic dialogue assembly, we hope to inspire future research in this promising area.

Chapter 2

Background

2.1 The Current State of Dialogue in the Field of Natural Language Processing

Because of the difficulty that individual components of dialogue present, tasks within the field of Natural Language Processing often focus on hyper-specific sub-components of dialogue. The phrase "dialogue in NLP" can refer to a myriad of different tasks such as the ability to generate entire conversations, generate a single response to a statement or entire conversation, and the development of chat bots that are able to converse on specific topics or on any subject. This is just a small glimpse into the range of tasks that comprise the study of dialogue.

A popular model within the world of chat bot development is the seq2seq model [10]. It is a model that takes in a sequence and outputs a sequence, which is well-suited to many text-processing tasks such as dialogue. Recent improvements in the seq2seq model training process, the model itself, and the data it trains on have resulted in several state-of-the-art (SOTA) chat bots such as BlenderBot 2.0 [7] [11] released by Meta in 2021. Models such as BlenderBot 2.0 often train on the exact sort of data that they will later be working with: dialogue data. But even SOTA seq2seq

chat bots struggle with consistency in long conversations and maintaining natural dialogue. While part of this is due to model constraints, it is also partially due to the fact that there is not much high-quality conversation data.

There are many notable dialogue datasets that have been extensively used across the past decade in developing dialogue models.

The Switchboard [4] dialogue dataset is a small repository of 2,400 phone conversations involving 70 topics. It is the highest quality dialogue data available: nearly perfectly natural conversation between humans. However, with only 2,400 conversations, it is not of sufficient quantity to train any large neural models.

DailyDialog [8] is a dataset of 13,118 conversations on topics related to daily life. It is good quality data, constructed from websites that help people practice English in everyday settings. Since the dialogue data was taken straight from websites, its quality is only as high as what the website provides, which is not always perfect or natural.

Topical-Chat [5] is a dataset of 10,784 conversations composed of statements within 8 topics: Fashion, Politics, Books, Sports, General Entertainment, Music, Sciences and Technology, Movies. It is good quality data, created by Amazon Turk workers. The method of creation however, also involved serving up related web information to the turkers on each turn of the conversation, and this often leads to awkward and artificial injections of facts into conversations that are not human-like.

PersonaChat [13] is a dataset of 10,907 conversations. It was created in the same way as Topical-Chat, using Amazon Turk workers, and similarly provided a 3-5 sentence "persona profile" that turkers were expected to adhere to when speaking. Because of this, it suffers from the same issue that Topical-Chat does, where facts or opinions are unexpectedly and strangely placed in parts of conversations where they don't fit naturally.

Wizard of Wikipedia [3] is a dataset of 22,311 conversations on 1,365 topics.

It is meant to simulate conversations between someone knowledgeable and someone curious, and was crowdsourced in a similar method to Topical-Chat and PersonaChat and so suffers from the same issues.

MultiWOZ [1] is dataset with 10,000 conversations among 7 service industry-related topics: Restaurant, Attraction, Hotel, Taxi, Train, Bus, Hospital, Police. It is a very consistent dataset, but the dialogues use formal language, making it less useful to most conversational chat bots.

Opensubtitles [9] is a dataset with millions of conversations taken from movie and television subtitle data. While the statements are natural, the actual data is relatively noisy due to the dataset's issues with properly segmenting and assigning speaker statements.

Perhaps the most important takeaway from all of these dialogue datasets is that none are high quality and also high quantity. Seq2seq models need tens of thousands of conversations to train on — though fewer are required for fine-tuning specifically — and being able to limit the topic of those conversations is necessary when training topic-specific bots.

2.2 Establishing Feasibility

The project began as a simple idea: what if conversations could be constructed from publicly available social media posts and their comments? Reddit, with millions of readily and openly available posts, was chosen as the best social media for this purpose. Alongside quantity, it has segmented topics due to the presence of subreddits, rich interactions because of its tree comment structure, and clearly defined posts which serve as sub-topics.

Other social media platforms didn't quite measure up in terms of features or availability. Facebook, for example, does not provide easy access to posts, and has

a chain comment structure which is less interaction-rich. While twitter does have an API for downloading posts, tweets are highly stylized and limited in length, and we wanted our conversations to be comprised of standard utterances. Additionally, tweets often contain images or videos which would disrupt the model.

What also made the project promising is that modern NLP models are quite powerful due to the extremely large amount of data that they are trained on. But while this gives them extensive capabilities in tasks such as next sentence prediction, it doesn't quite render them fluent enough to generate quality dialogue. Our model takes advantage of these models' newfound strengths to sequence existing human-written content into natural dialogue while avoiding the two difficult tasks of generating natural and relevant responses.

2.3 Initial Ideas

During the initial phase of the project, we developed a number of different potential approaches. Many of these approaches were outside of the scope of the thesis, but they are included here as possible future work and as a representation of how we were initially conceptualizing the project: an amalgamation of modern NLP models.

The first method we considered was simple: taking the entirety of a post to be the first statement in a dialogue, and then using top-level comments as potential replies to the post. The second method made use of Reddit's feature where a user can respond to a specific part of the original post. We considered taking the parts of the original post and matching them to the comments that responded to them and treating these as dialogues. Both of these methods would provide extremely high-quality dialogue, but clearly neither provide multi-turn dialogue.

The third method would map comments to sentences of the original post. We would split the original post into individual sentences and then use a language model

or dialogue ranker to decide which comment is most likely to follow each individual sentence. This would allow for a multi-turn dialogue, although Speaker 1 would be unable to respond to Speaker 2. This method became the core of the model we developed.

The fourth method proposed using a pre-trained Question-Answering model to try to answer questions that were asked in the original post by finding the answer among the comments. This would serve as a single interaction in a longer dialogue created by previous methods.

The fifth method considered the possibility of using a state-of-the-art chat bot such as BlenderBot 2.0 [7] [11] to smooth over interactions in the dialogue that we determined to be low-quality by either generating an entire substitute statement or generating part of a statement to attach to the low-quality interaction. This method was investigated, and later discarded because even as a SOTA bot it often provided statements that were worse than the original we were aiming to replace.

The sixth method looked at using summarizing models to shorten posts and comments to make them more suitable lengths for dialogue interactions.

The seventh method proposed making a language model generate a response to a sentence of the original post and then using the same model to identify which of the post's comments was most similar to the generated response.

The eighth, ninth, and tenth methods were all concerned with maintaining each individual speaker's style throughout the conversation. For example, using diction, syntactical, or sentiment analysis to keep the chosen statements relatively consistent with previous statements.

Of this list, the third and fifth methods were deepest investigated; the third method, with several modifications, provided the best results.

2.4 Beneficial Reddit Properties

In addition to the basic features mentioned earlier, Reddit has a couple other aspects that can be used when assembling dialogue.

For example, some subreddits have specific tags that users must employ in order to comment on a post; these tags include labels such as "response to question" and "further question". As well, a single comment tree can be already considered a conversation of sorts, in which consecutive replies likely deal with the same topic; while we could easily source many conversations from simply taking comment chains, the resulting model would not be extensible or modifiable, and additionally such conversations would lack context and content.

Reddit also has a variety of styles of content. In some subreddits, all of the posts are questions and the comments are answers, where in other subreddits all of the posts are stories and the comments are reactions. There are even some subreddits, such as r/casualconversation, which are centered around having conversations.

All of these aspects provide even more for models to work with when assembling dialogues.

2.5 Connections to the Emory Natural Language Processing Lab

This project was designed in relation to several ongoing projects in the Emory Natural Language Processing Lab. Sophy Huang, a concurrent honors student, is working on a project in detecting emotion patterns in dialogues; this is a project with clear applications for our model, which could make use of such patterns to better sequence statements. James Fillwock and Sarah Fillwock are working on Emora, a general-domain chat bot, which could make use of dialogue data we generate to train the

various models it uses. There are additionally projects related to domain-specific chat bots, such as a bot that can speak with students about college-related issues that would make use of topic-specific dialogue data.

2.6 Collaboration

Early on in the project, we realized that the amount of work required to fully explore all of the avenues we had imagined would require more than just one student working alone. Because of this, with Dr. Choi's advice, the project was a collaboration with another honors student, Daniil Huryn, who worked more closely with the neural aspect of the project, undertaking fine-tuning models, training seq2seq models, setting up GRADE [6] (a post-filtering metric), and working with BlenderBot 2.0 [7] [11] during the process, as well as assisting with the core of the final approach. His solo work will not be referenced in this thesis, though the final results of the model, of course, are shared.

2.7 Beginnings

The model began as a simple combination of method three and five, with the intent to use both next sentence prediction and BlenderBot 2.0 in order to create natural conversations.

The data that we developed the model with was 36,044 Reddit posts from a collection of eight college-related subreddits scraped by Dr. Choi: r/ApplyingToCollege, r/AskAcademia, r/College, r/CollegeAdvice, r/CollegeMajors, r/CollegeRant, r/Emory, and r/GradSchool. These subreddits provided a good representation of the different forms that Reddit data often takes: questions, advice, rants, stories, and opinions.

Chapter 3

Evaluation Techniques

3.1 Automatic Dialogue Metrics

Dialogue metrics are useful for two simple reasons: we can use them to determine which of our model's variants performs the best, and we can also use them to pick which dialogues we want to keep based on their quality.

Ideally, this evaluation would be done automatically without human intervention. Unfortunately, due to the same difficulties discussed earlier, it is quite difficult to do so. However, there do exist several automatic dialogue assessment metrics, each with various strengths and weaknesses [12].

For our project, we chose to take the average next sentence prediction score and an automatic metric called GRADE [6] in order to filter out bad dialogues that we had created.

3.2 Rubric for Manual Evaluation of Generated Dialogues

Because automatic dialogue metrics are yet imperfect, we developed a rubric for manually annotating conversations that we could use while developing our model. This allowed us to choose between variations of the model as well as select values for models' hyper-parameters.

The first rubric that we used, see Table 3.1, had a couple issues, foremost of which was that scores 2-5 were occupied by models that came close to achieving realistic dialogue. Because of this, a model with an average score of 2 was very different in quality than a model with an average score of 1. Effectively, we weren't doing a great job at showing differences in quality between models because of how our rubric misrepresented the model space.

Table 3.1: Initial Rubric

Score	Description
1	Nonsensical
2	Mostly Nonsensical
3	Coherent but not humanlike
4	Only a few issues
5	Perfect

Table 3.2: Final Rubric

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

Because of the issues we found with the initial rubric, we decided to design a better rubric that would grade conversations in an intuitive way while properly separating

models of different quality across the entirety of our model space.

The final rubric, see Table 3.2, avoided the misrepresentation of the first rubric. The rubric design scales score with the number of good interactions. This is because our data is only useful to models in proportion to the amount of good interactions it provides to train and test on. Since the manual grading of dialogues must be subjective, we chose a definition of "good interaction" that would make annotating conversations as closely aligned with human dialogue-related intuition as possible: "Is this a normal response or continuation of the previous statement?".

The weakness of this rubric, however, is that it doesn't represent the quality of the data for a model trained on multiple interactions in a row. For example, we didn't grade on consistency across statements as this was lower priority to us than the naturalness of the conversation.

Chapter 4

Development of Model

4.1 Initial Control Flow

The core of our model is the control flow, or the basic idea of how we use the post and comments to construct a dialogue. For our control flow, we chose to take the original post's individual sentences as a guide for the Speaker 1 component of the dialogue, using comments to fill in Speaker 2's responses. While this method has a few clear flaws, they are lessened by the modifications discussed later in this chapter. The algorithmic approach is described in Algorithm 1 below.

Using the original post as a boundary for Speaker 1's statements allows for our dialogue to remain a relatively compact semantic unit. There is always one story or continued statement taking place across the dialogue that eventually comes to a resolution near the end. This helps smooth out any awkwardness from beginning or ending dialogues, while keeping the conversation centered around one semantic path.

We also decided to combine the title of the post with the first sentence so that they would also be seen in conjunction, as from our analysis of posts this often made sense and rarely hurt the dialogue.

An example of how a post and comments might be turned into a dialogue using this

control flow is given below in Figure 4.1. This post and its comments are simplified from the example Reddit post and comments in Appendix E.

Algorithm 1: Base Approach

Input : Post Text, Comments

Output: Conversation

```
current_conversation ← set to empty list;
repeat
    Append next sentence from post to current_conversation;
    comment_scores ← set to empty list;
    repeat
        comment ← next comment in input comments;
        comment_score ← NSP score of (current_conversation, comment);
        Append comment_score to comment_scores;
    until No more comments left;
    Take comment associated with highest score in comment_scores and
     append to current_conversation;
    Remove chosen comment from available comment choices;
until No more sentences left in post;
```

4.2 Next Sentence Prediction: BERT vs DialogRPT

To choose comments for Speaker 2 to use to respond to Speaker 1, we needed a model that could take in context and choose from a list of possible responses or continuations to that context. Two types of models, dialogue rankers and next sentence prediction models, fit this need. We tried one of each when beginning this project, BERT's NSP [2] head and DialogRPT [14].

Dialogue rankers are designed to choose continuations of dialogues, while NSP

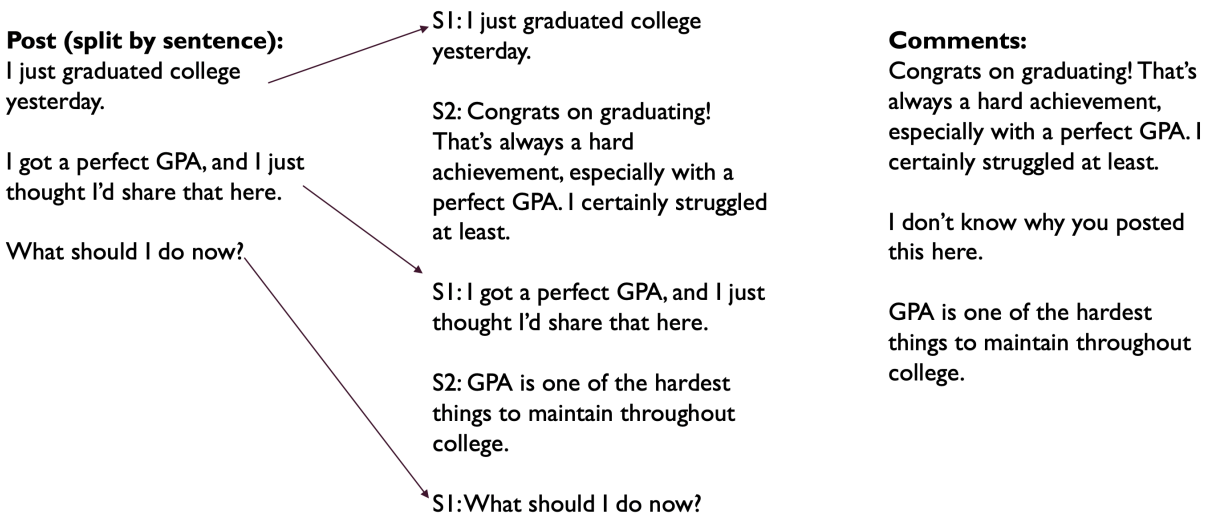

Figure 4.1: Basic Control Flow Example

models are only optimized to choose a next sentence, so we expected DialogRPT to easily outperform BERT, however, we discovered that this was not the case. In our manual evaluation of dialogues generated using the above method and then BERT NSP and DialogRPT as the Speaker 2 response chooser, we found that BERT NSP significantly outperformed DialogRPT.

Table 4.1 shows the extent to which BERT NSP was found to outperform DialogRPT, with a 1.38 increase in score based on 50 manually-evaluated generated conversations for each. Even though BERT NSP tended to generate shorter conversations, this was less important to us than the 20x increase in speed and nearly 2x score performance.

Table 4.1: Comparing BERT NSP and DialogRPT

Model	Turns per Second	Score	Score Std. Dev.	Turns per Convo.
DialogRPT	0.16	1.53	1.07	4.57
BERT NSP	2.02	2.95	1.43	2.45

Because of this, we chose to use BERT NSP as the base of our model.

4.3 Initial Generated Data and Error Analysis

The initial generated data from this method was promising but had several issues. By far the biggest issue of this first approach was that Speaker 1 was unable to respond to comments that Speaker 2 made. This led to one-sided conversations. In addition, because Speaker 1's statements were all single sentences, aside from the first statement, the conversation felt stilted. Even worse, at this point, all comments were included in their entirety, when often only part of a comment was relevant.

Below, in Table 4.2 is a short excerpt from a conversation generated on a post from r/CollegeRant by this early approach. Speaker 2's responses are quite relevant to what Speaker 1 says, but Speaker 1 does not address what Speaker 2 says. While this sort of dialogue is not ideal, it is not far from realistic dialogue. Note that at this point in the model, sentence boundaries were determined by splitting on periods, and because of this the first utterance that Speaker 1 has in the conversation is two sentences.

Table 4.2: Initial Method Example Dialogue

Speaker	Utterance
Speaker 1	"Why do classes give out so much work? I'm not just talking about moving to online, either"
Speaker 2	"I'm not even working rn because of my schoolwork amount"
Speaker 1	"Ever since my very first semester at college, my professors have been pilling on the reading homework and quizzes and assignments back to back to back"
Speaker 2	"I can relate to this as the quizzes and questions are so heavy it's like the point is just do get them done not even use ur critical thinking skills"

4.4 Continuing Development

The initial method showed enough promise that we decided to continue developing with this method as the core of the model. As our next features, we wanted to add

multi-sentence Speaker 1 statements; find a way to allow Speaker 1 to respond, or at least pseudo-respond, to Speaker 2's comments; and we wanted to allow partial comments to become Speaker 2 statements in case the entire comment was not relevant to Speaker 1's utterance.

4.5 Advanced Control Flow

The first modification we made to the initial model was allowing the model to choose the next sentence in the original post instead of a comment from the post; if it chose the next sentence in the original post, then that sentence was combined with the previous and the method would start over. This allowed for Speaker 1 to have multi-sentence utterances and greatly improved the naturalness of their statements. This approach is shown in Algorithm 2.

Algorithm 2: Improved Approach

Input : Post Text, Comments

Output: Conversation

```
1  current_conversation ← set to empty list;
2  repeat
3      Append next sentence from post to current_conversation;
4      option_scores ← set to empty list;
5      repeat
6          comment ← next comment in input comments;
7          comment_score ← NSP score of (current_conversation, comment);
8          Append comment_score to option_scores;
9      until No more comments left;
10     Append NSP score of (current_conversation, next sentence from post) to
        option_scores;
11     Find highest-rated option in option_scores;
12     IF a comment, then append to current_conversation, and return to line 3;
13     ELSE add to the last entry in current_conversation, and return to line 4
14 until No more sentences left in post;
```

Then, we pre-processed top-level comments by taking series of sentences up to size 3 across each comment as well as including the entire original comment, as shown in Algorithm 3. We chose not to segment lower-level comments for computational complexity reasons primarily, though additionally lower-level comments were less likely to be relevant to the original post. Segmenting top-level comments helped because often comments would be quite large and address individual parts of a post in separate sections.

Algorithm 3: Comment Segmentation

Input : Post Top-Level Comments

Output: Varied-size Comment Segments

```
comments ← set to top-level comments in post;
final_comments ← set to comments;
repeat
    comment_sentence_list ← list of comment's sentences;
    counter ← set to size 1;
    repeat
        segment_size ← set to 1;
        repeat
            Append entries at indices 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;
return set of final_comments to avoid duplicate entries
```

4.6 Beam Search

At this point, we realized that beam search would be a great addition to our model. Beam search works well in models that are trying to moderate the effect that greedy local decisions have on the overall quality of the output.

Beam search was not implemented to allow for paralellization, because at the time that it was implemented, we were exploring other model options that relied on a single state throughout. Instead, a list of conversation states was maintained and updated

one at a time, illustrated below in Algorithm 4.

Algorithm 4: Beam Search Overview

Input : Post, Beam Search Size n

Output: Conversation

1 *conversation_states_list* ← set to first sentence of post;

2 **repeat**

3 | generate n responses to n current states;

4 | *conversation_states_list* ← top n conversation states by NSP average;

5 **until** *each state is done generating dialogues*;

From just these three improvements, we saw a drastic improvement in score, as shown in Table 4.3. Compared to the initial method's score of 2.95, our new method without beam search showed an improvement of 0.82 points. With even just size 2 beam search added on, this score improved another 0.36 points. The turns generated per second, as expected, fell linearly as we increased the beam size.

Table 4.3: BERT NSP Beam Search Performance

Beam Size	Turns per Second	Score	Score Std. Dev.	Turns per Convo.
1	0.48	3.77	0.971	10.01
2	0.33	4.125	0.77	10.08
4	0.12	4.159	0.84	10.58
8	0.06	4.125	0.74	10.01

Due to time complexity constraints and minimal improvement after beam size of 2, we decided to stick to beam size 2 for our model.

At this point conversations were much more natural. An example excerpt is given below in Table 4.4. Now, Speaker 2 responses are able to be partial comments and Speaker 1 statements exist as more compelling semantic segments.

Table 4.4: Beam Search BERT NSP Method Example Dialogue

Speaker	Utterance
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."

4.7 Threading

However, even with the previous improvements, Speaker 1 was still not addressing anything said by Speaker 2. To attempt a fix at this issue, we used a method we termed Threading.

When we take a comment from the comment section for Speaker 2's response, there are often comments that reply directly to it. We used segments of these comments, when available, as prefixes to Speaker 1's following statements. As shown in Algorithm 5, we simply took the current conversation state and the comment we'd chosen to serve as Speaker 2's response and then checked whether using one of its comment responses would improve our score.

Relevant here is that because the threading option is from a different origin than Speaker 1, whose content is drawn from the original post, we penalized the score of the option based on its length, i.e. its number of tokens. This was done via the equation $score = NSP(prompt, option) * (1.15 - (0.01 * option\_length))$. This helped us avoid very long responses to Speaker 2, which have high potential to clash with the rest of the statement made by Speaker 1.

Algorithm 5: Threading

Input : Current Conversation State, Comment Chosen

Output: Modified Conversation State with Potential Threading Response

```
threading_options ← list of comment responses to input CommentChosen;
repeat
    option_scores ← set to list containing score of Input Comment Chosen;
    repeat
        threading_option ← next comment response from threading_options;
        option_score ← NSP score of (current_conversation + Input
         CommentChosen, threading_option);
        Append option_score to option_scores;
    until No more comments left;
    Find highest-rated option in option_scores;
    IF a comment, then append to current_conversation;
    ELSE, continue with normal conversation creation;
until No more comment responses left;
```

Now, Speaker 1 no longer ignored Speaker 2's utterances, partially solving a long-standing issue with our approach. The flaw was not entirely resolved as not every comment has a comment that responds to it, but around 25% of the time we were able to use the threading to augment statements made by Speaker 1.

This led to interactions such as shown in Table 4.5. Additions made by threading have been emphasized. In this excerpt, threading provides natural responses without interfering with the rest of the statement. This modification took further advantage of the rich interactions that Reddit has, which is what made it so strong.

An illustration of the final advanced control flow is shown in Figure 4.2. Threading's contribution to the final dialogue is shown by bold and italicized text.

Adding threading on top of the BERT NSP beam search model caused our score

Table 4.5: Threading Method Example Dialogue

Speaker	Utterance
Speaker 2	"Doesn't it cut both ways? You say we made racist jokes. Just tell him that if this happens you'll send his jokes to his profs."
Speaker 1	"*Nope.* I'm honestly super scared because I don't know what would happen if the professor saw the screenshots"
Speaker 2	"maybe don't make racist jokes in the first place?? actions have consequences :)."
Speaker 1	*Yes, I recognize that.* My friend knows my real life name, but the screenshots don't include my full name"

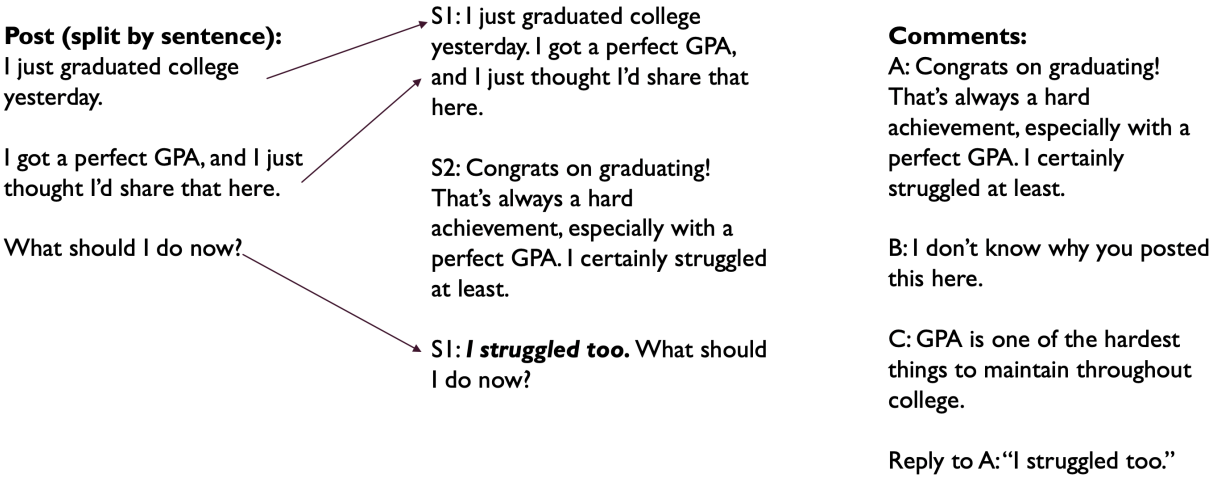

Figure 4.2: Advanced Control Flow Example

to raise to 4.26 from the previous score of 4.125. This served as our final approach and it produced dialogue at a speed of 0.05 turns per second. Though this speed was much slower than our original approach, by running multiple tasks it was relatively easy to generate upwards of 1000 conversations an hour.

We ran our model on 1190 posts from r/books and 770 from r/movies, 1960 total posts, and ended up with 951 dialogues, which is a conversion rate of 48.52%. For our model, not every post is converted into a dialogue, and not every generated dialogue is kept. Around 40% of posts were never converted into a dialogue. This was because a lot of posts didn't have any comments or the posts themselves were too short, sometimes consisting of only a title; we also removed posts that had links or certain punctuation, listed in Appendix D, inside of them. After generating all of the posts, we removed the bottom 5% of posts by average NSP score, then we removed 10% based on the automatic dialogue metric GRADE.

Chapter 5

Evaluation of Final Model

5.1 Evaluation Set-up

Beyond manual annotation by the dataset's developers, evaluation of dialogue datasets can take several forms. Perhaps the most direct measure of a dataset's use is how they contribute to a model's improvement in performance. This would entail training several models on our generated dialogues and comparing the resulting models to those trained on other datasets. This method, unfortunately, was outside the scope of this thesis.

Fortunately, a much simpler and perhaps even more useful evaluation exists in the form of direct comparison of our conversations against popular dialogue datasets.

5.2 Comparison Datasets

For our target comparison datasets, we wanted to choose good/high-quality datasets that were of medium size ($\sim$10k conversations) and had been created in varied styles. For us, if we demonstrated higher quality data, or even just the same level of quality as these datasets, then with our ability to arbitrarily generate millions of dialogues, we would have an important and valuable result.

Therefore we chose Topical-Chat [5], DailyDialogue [8], and MultiWOZ [1]. In Tables 5.1 and 5.2 the statistics and topics of these datasets against our dataset, named HuHu, are displayed. Our dataset had around the same number of turns per conversation as MultiWOZ and HuHu on average, though it was far behind Topical-Chat. For average Length of Utterance, our dataset led by over 7 tokens per utterance; as will be discussed later, this was part of the reason our conversations appeared more human-like and casual than other datasets. Though we only generated 951 chats, this is relatively unimportant as we are able to generate as many conversations as needed. We chose books and movies as topics as a way to match some of the content we'd be comparing against in Topical-Chat and because they are common topics of conversation.

Table 5.1: Comparison Datasets' Metrics

Dataset	Turns per Convo	Length of Utterance	# of Chats
Topical-Chat	21.8	19.6	10,784
MultiWOZ	13.46	13.13	10,000
DailyDialog	8	15	13,118
HuHu	8.88	26.63	951

Table 5.2: Comparison Datasets' Topics

Dataset	Topics
Topical-Chat	Fashion, Politics, Books, Sports, General Entertainment, Music, Sciences and Technology, Movies
MultiWOZ	Restaurant, Attraction, Hotel, Taxi, Train, Bus, Hospital, Police
DailyDialog	Daily Life
HuHu	Books, Movies

5.3 Amazon Mechanical Turk Task Design

To compare against these datasets, we initially used the Amazon Mechanical Turk platform to crowd-source evaluation.

The task had a couple original possible designs: we could provide turker's with our rubric and ask them to grade individual conversations and then use these scores to rank datasets based on the average score of their dialogues or we could provide turker's with two dialogues and ask them which is better or ask them to rate how much better or worse on a scale each is.

We ended up deciding on providing two dialogues, one from our dataset and another from a target comparison dataset, and asking turker's to rate the naturalness of one conversation to another on a scale of 1 to 5. The task interface is presented in Figure 5.1 below.

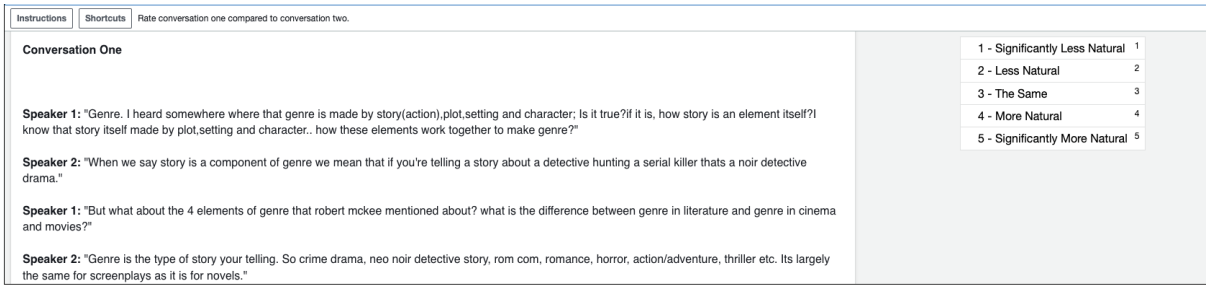

Figure 5.1: Amazon Turk Task Interface

Conversation One and Two were displayed in order, randomized between our conversation and the dialogue from the other dataset, and the instructions were simply, "Rate conversation one compared to conversation two.". If further instruction was needed, there was an additional instruction section which said, 'Rate conversation one compared to conversation two. If conversation one reads much more naturally than conversation two, then select significantly more natural. If they feel the same, then pick "The Same", and if conversation two is much more natural, select "Significantly Less Natural"'.

To avoid potential bias in grading, we eliminated the degrees of freedom between dialogues being compared. All comparisons were done on dialogues with the same number of turns and closest to the same average utterance length. This was not possible for Topical-Chat due to the fact that its shortest dialogue was 20 turns,

while our longest was 17, so all Topical-Chat comparisons took place with dialogues that were 20 or 21 turns long, i.e. the shortest possible, while finding the closest average utterance length.

Turker's were given 3 minutes to compare two conversations and were compensated at a rate of $6/hour. Our evaluation consisted of 400 conversations across two topics, books and movies, compared against 400 conversations taken from Topical-Chat, MultiWOZ, and DailyDialog combined. Each comparison was completed by two separate turker's to determine trustworthiness of the results.

5.4 Results

5.4.1 Disproving Turk Results

The results were very positive for our dialogues; unfortunately, we realized that Turkers tended to just rate whichever conversation came first more positively. Below, in Table 5.3 and Figures 5.2 and 5.3 the results from the crowdsourcing are shown. The results are positive because our turk task did not have a good split between placing our conversations first or second; our conversations were placed first in 70% of tasks, skewing the results toward our dialogues.

Table 5.3: Overall Results (Total 800)

Score	Number	Percentage
1 - Significantly Less Natural	32	4%
2 - Less Natural	134	16.75%
3 - The Same	181	22.63%
4 - More Natural	351	43.88%
5 - Significantly More Natural	102	12.75%

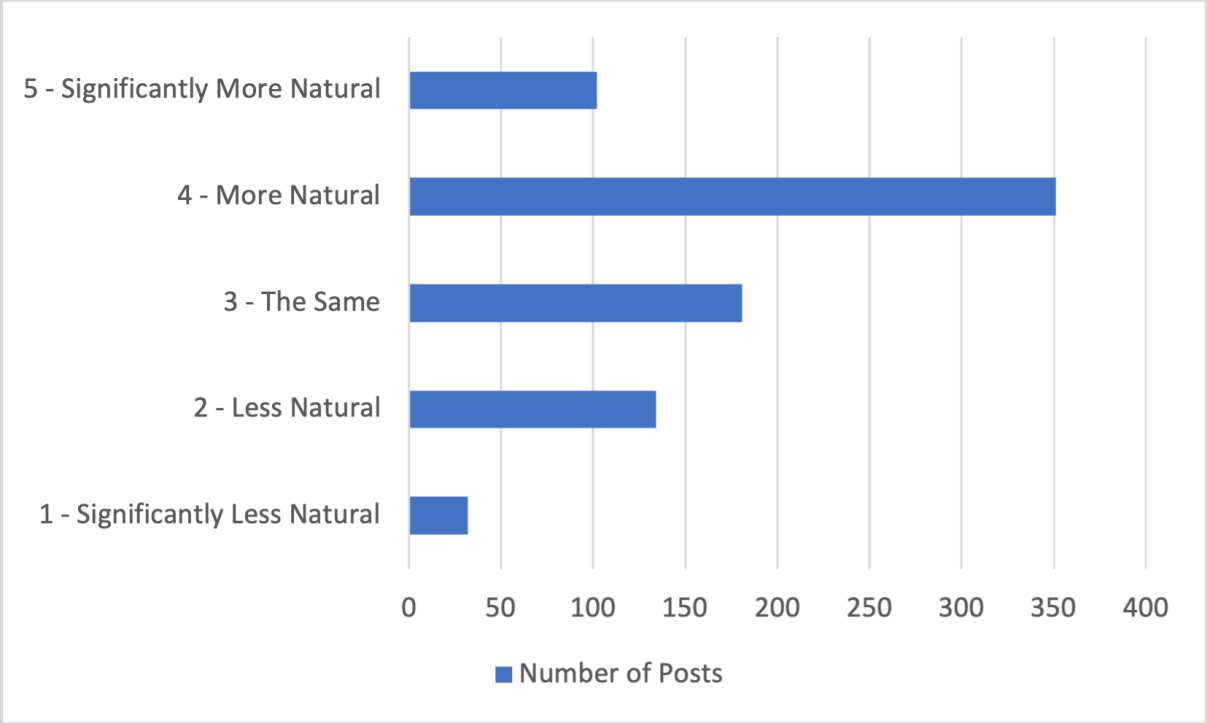

Figure 5.2: Overall Scores Distribution

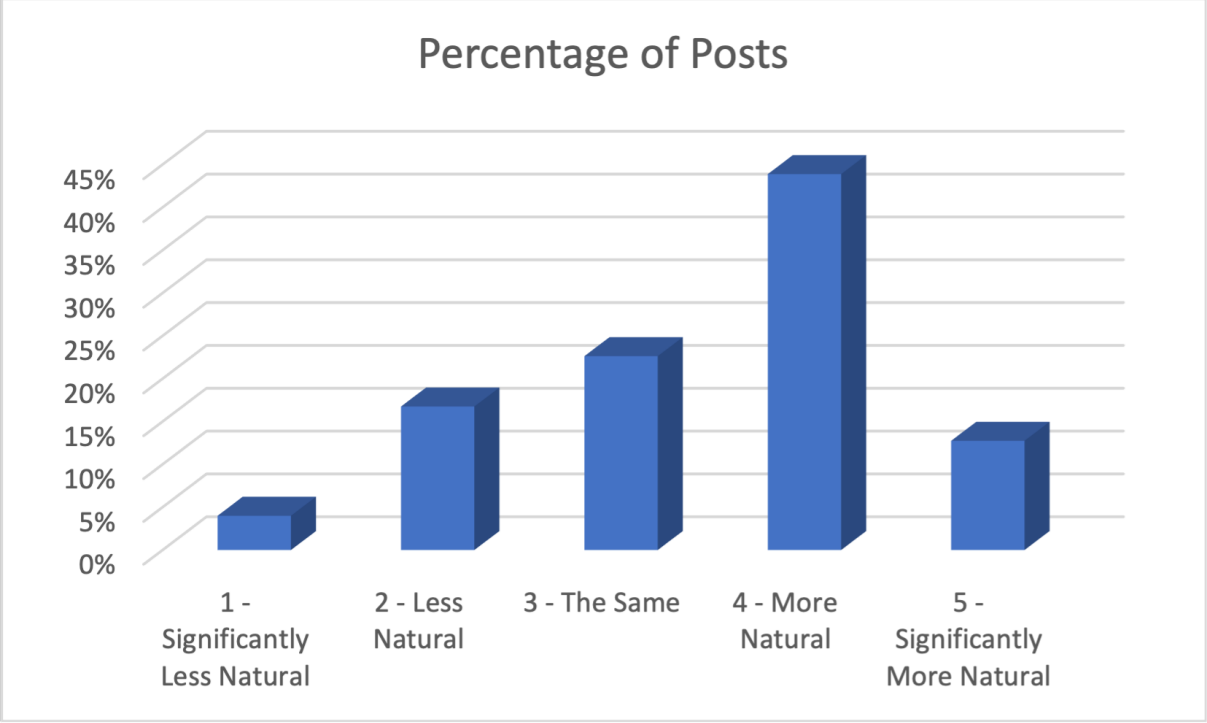

Figure 5.3: Overall Scores Percentage Distribution

Using Amazon Turk presents a few risks in bringing in workers who sometimes auto-select a random answer and submit quickly, therefore we also checked the results based on the amount of time the worker took to complete the task to see if this would help the randomness. Unfortunately, as shown in Tables 5.4 and 5.5, results did not change or become balanced as workers took more time.

Table 5.4: Results for >30 Seconds (Total 639)

Score	Number	Percentage
1 - Significantly Less Natural	26	4.1%
2 - Less Natural	105	16.43%
3 - The Same	148	23.16%
4 - More Natural	280	43.82%
5 - Significantly More Natural	80	12.52%

Table 5.5: Results for >60 Seconds (Total 400)

Score	Number	Percentage
1 - Significantly Less Natural	21	5.25%
2 - Less Natural	63	15.75%
3 - The Same	87	21.75%
4 - More Natural	174	43.5%
5 - Significantly More Natural	55	13.75%

Another good proof of the randomness of the results is that our dialogues scored equally across all three datasets and both of our topics even though we had personally verified that not all of the datasets were the same level of quality. The lack of change in results is shown in Tables 5.6, 5.7, 5.8, 5.9, and 5.10.

Table 5.6: Results for >60 Seconds against DailyDialog (Total 174)

Score	Number	Percentage
1 - Significantly Less Natural	13	7.47%
2 - Less Natural	25	14.37%
3 - The Same	36	20.69%
4 - More Natural	78	44.83%
5 - Significantly More Natural	22	12.64%

Table 5.7: Results for >60 Seconds against Topical-Chat (Total 134)

Score	Number	Percentage
1 - Significantly Less Natural	5	3.73%
2 - Less Natural	23	17.16%
3 - The Same	32	23.88%
4 - More Natural	53	39.55%
5 - Significantly More Natural	21	15.67%

Table 5.8: Results for >60 Seconds against MultiWOZ (Total 92)

Score	Number	Percentage
1 - Significantly Less Natural	3	3.26%
2 - Less Natural	15	16.30%
3 - The Same	19	20.65%
4 - More Natural	43	46.74%
5 - Significantly More Natural	12	13.04%

Table 5.9: Results for >60 Seconds Books Topic (Total 253)

Score	Number	Percentage
1 - Significantly Less Natural	11	4.35%
2 - Less Natural	45	17.79%
3 - The Same	52	20.55%
4 - More Natural	104	41.11%
5 - Significantly More Natural	41	16.21%

Table 5.10: Results for >60 Seconds Movies Topic (Total 147)

Score	Number	Percentage
1 - Significantly Less Natural	10	6.81%
2 - Less Natural	18	12.24%
3 - The Same	35	23.81%
4 - More Natural	70	47.62%
5 - Significantly More Natural	14	9.52%

Further analysis on randomness of Turk results is presented in Appendix F.

5.4.2 Manual Annotation Results

Because we could not use the crowdsourced results, we manually double-annotated 100 dialogue comparisons ourselves, applying the same principles of removing degrees of freedom between compared conversations before starting. The resulting data is shown in Figure 5.4 below. The average score was 3.21; we were at least as natural as the other conversation in 73% of conversation comparisons, beating the other conversation in 46% of the tasks.

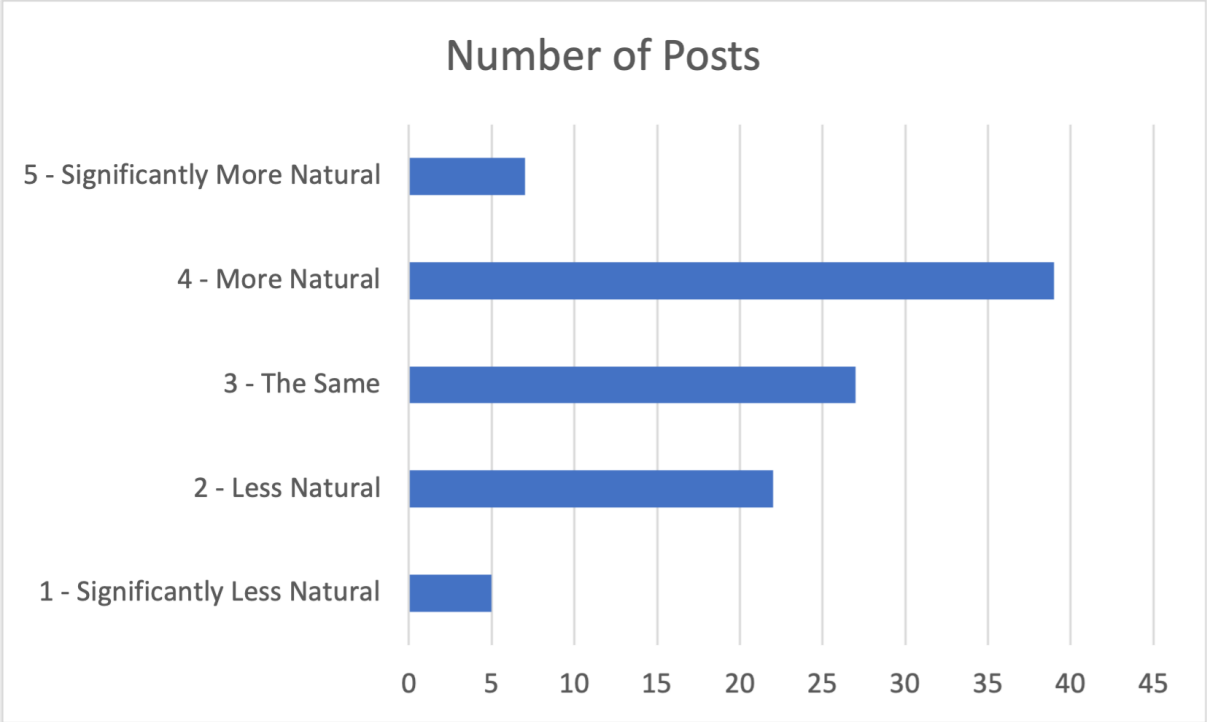

Figure 5.4: Overall Scores from Manual Annotation

The breakdown by database is shown in Figures 5.5, 5.6, and 5.7. By far, we outscored Topical-Chat the most frequently; DailyDialog and MultiWOZ were both much stronger than Topical-Chat, and we scored very slightly better than both. The exact percentages are shown in Table 5.11. For Topical-Chat, we scored at least the

same in 85.19% of comparisons and better in 70.37%; for MultiWOZ, we scored at least the same in 76.93% of tasks and better in 34.62%; for DailyDialog, we scored at least the same in 64.87% of comparisons and better in 37.84%. Note that these scores were calculated using only ∼33 comparisons per database, so our sample size is small.

Table 5.11: Manual Annotation Results

Score	Topical-Chat	MultiWOZ	DailyDialog
1 - Significantly Less Natural	0%	11.54%	0%
2 - Less Natural	14.82%	11.54%	35.14%
3 - The Same	14.82%	42.31%	27.03%
4 - More Natural	51.85%	30.77%	35.14%
5 - Significantly More Natural	18.52%	3.85%	2.70%

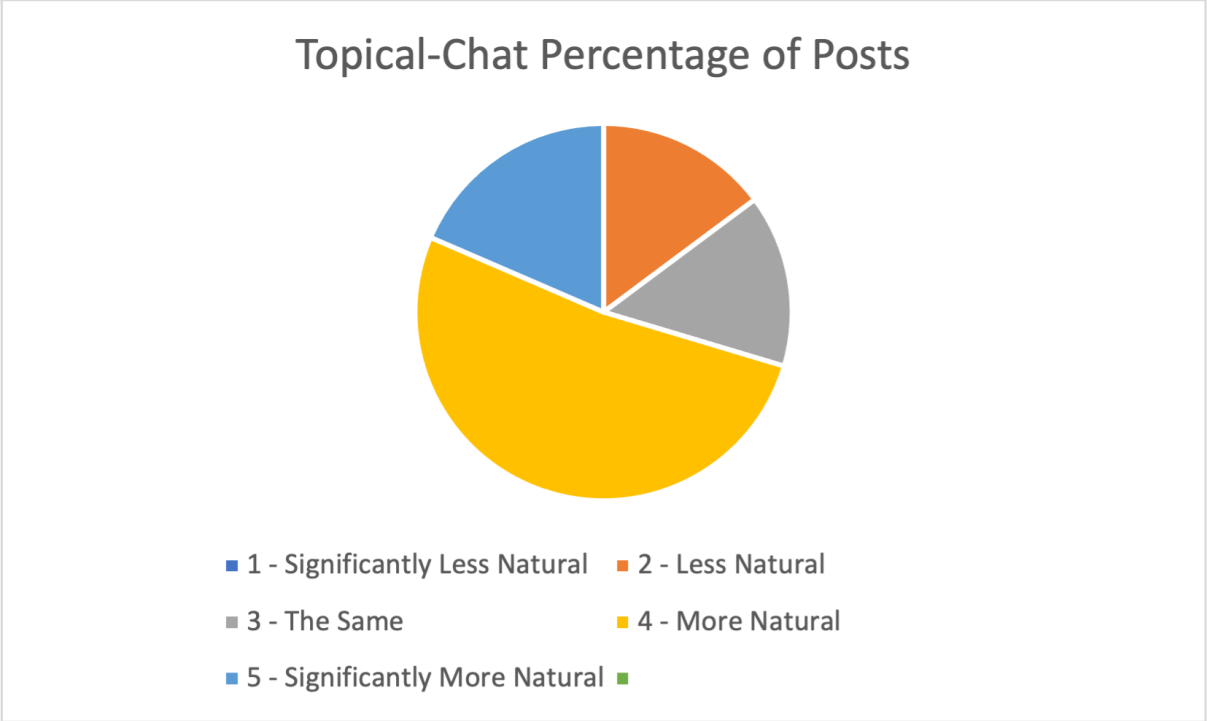

Figure 5.5: Topical-Chat Visual Distribution of Scores

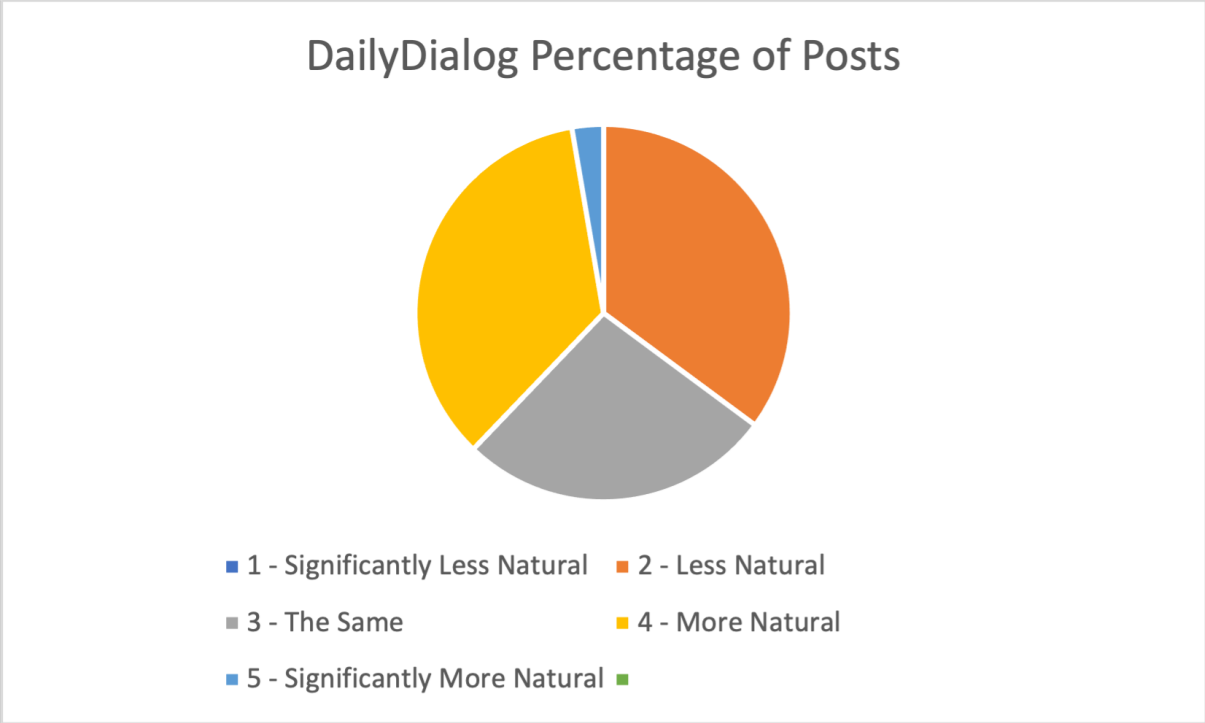

Figure 5.6: DailyDialog Visual Distribution of Scores

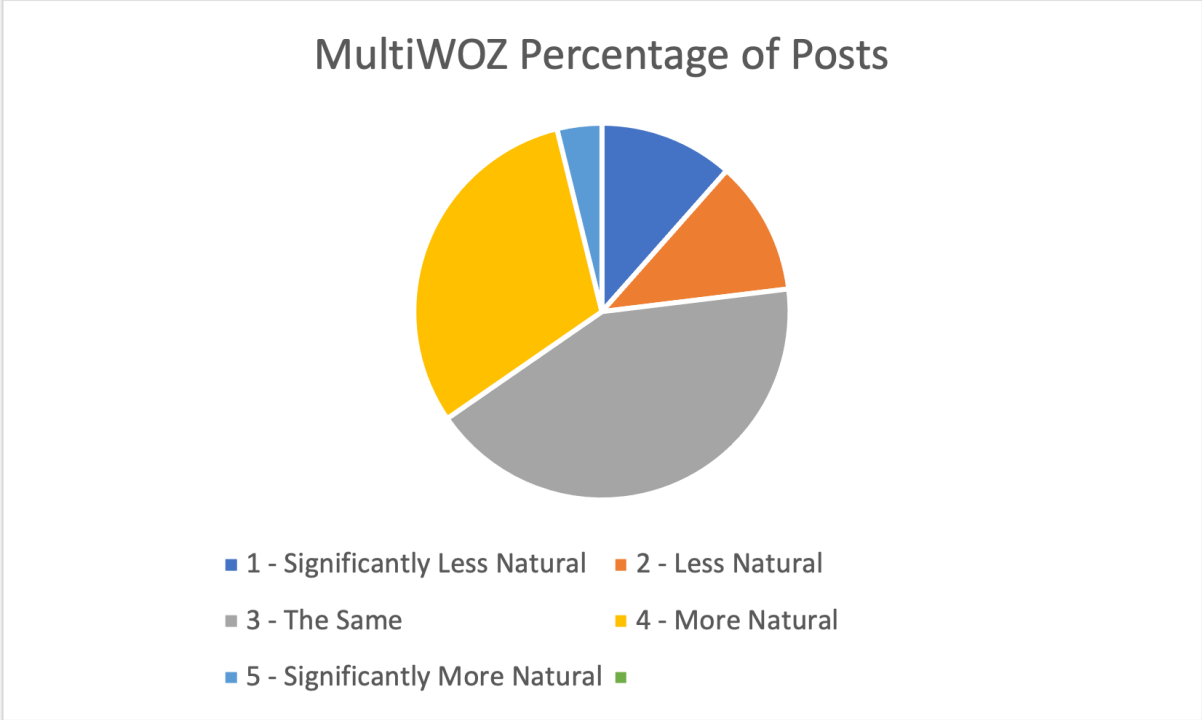

Figure 5.7: MultiWOZ Visual Distribution of Scores

As with the Turk evaluation, we broke down our scores by topic. This is visually represented in Figures 5.8 and 5.9 and shown in Table 5.12. What we found is that dialogues produced from posts taken from r/books tended to produce slightly worse conversations though it had more standout conversations than r/movies. Numerically, r/books posts led to dialogues of at least the same quality in 78% of cases, better in 46%, while r/movies led to dialogues of at least the same quality in 68% of cases, better in 46%. Note that these results are from a sample size of 50 comparisons each.

Table 5.12: Manual Annotation Results By Topic

Score	Movies	Books
1 - Significantly Less Natural	6%	4%
2 - Less Natural	16%	28%
3 - The Same	32%	22%
4 - More Natural	44%	34%
5 - Significantly More Natural	2%	12%

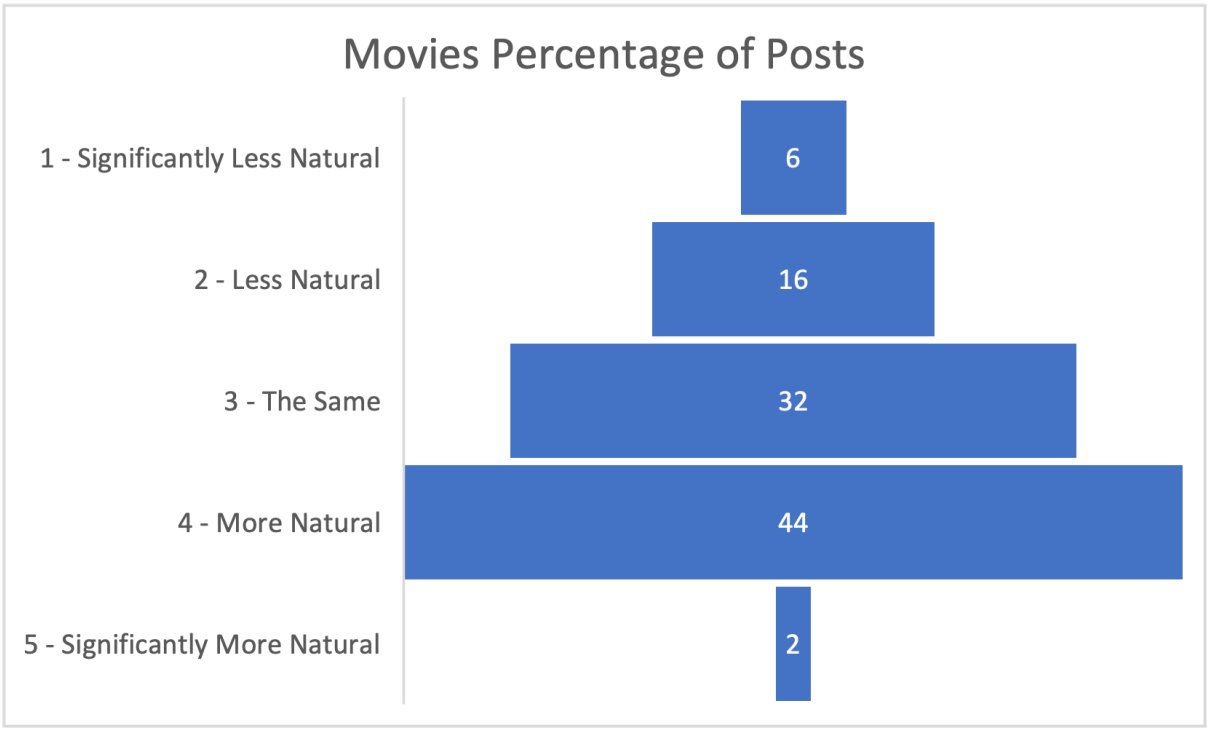

Figure 5.8: Movies Visual Distribution of Scores

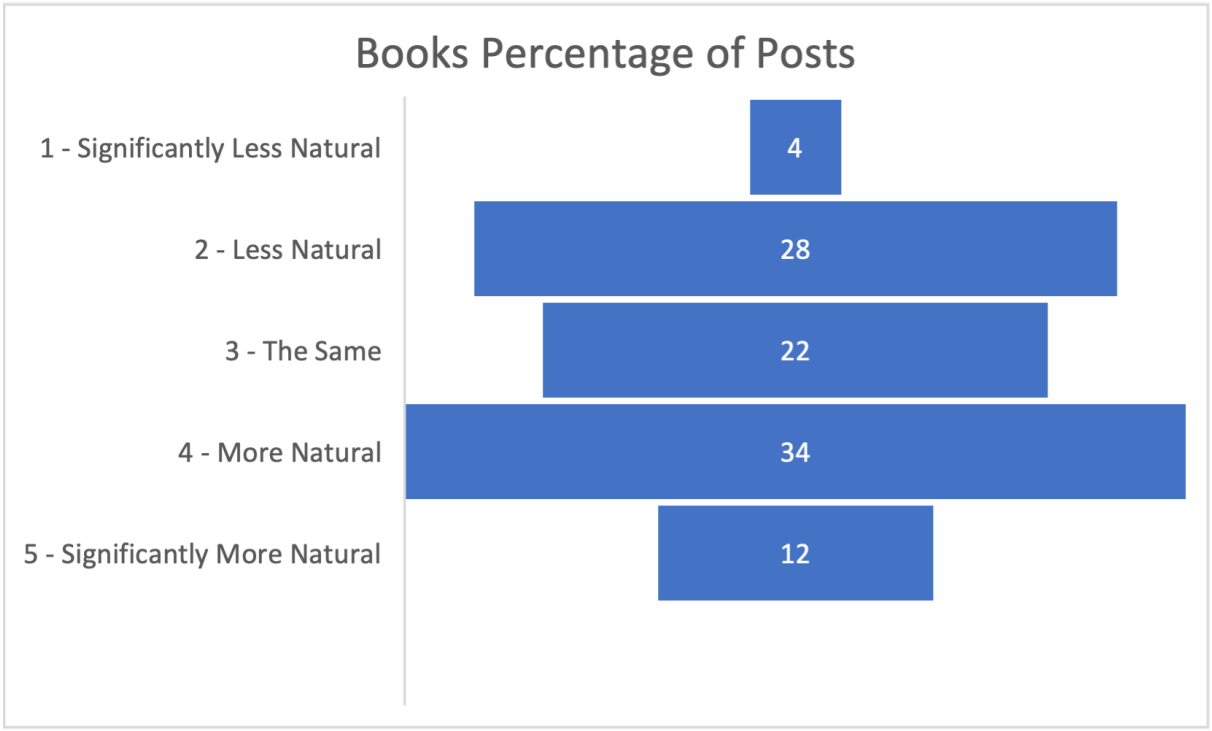

Figure 5.9: Books Visual Distribution of Scores

Chapter 6

Discussion

6.1 Model Analysis

6.1.1 Model Strengths

The strength of our final model is predicated on the strength of Reddit's interactions. While the use of the BERT NSP head is important, it isn't powerful enough to construct quality dialogues unless the base material is strong. This is a lesson we learned in the exploration of what we would include in our final model. Several other approaches that we looked at, including BlenderBot 2.0 smoothing and seq2seq reactions, didn't ever work perfectly because we were trying to generate language at the same level of quality as actually natural and relevant human language. For this reason, the work that we did around control flow, comment handling, and threading became particularly important to the success of our model.

6.1.2 Current Flaws

While our model produces dialogue data at a level of quality high enough to be useful, it still has a couple flaws that lead to poor output dialogue at times. Perhaps the biggest issue is that our model focuses on fluid conversation, not intelligent, sensible,

or helpful conversation. We have no way to account for truth or context, both of which are highly important for dialogues which will be used for training down the line. Additionally, because of the lack of awareness of context, our speakers sometimes have low consistency between statements.

Another issue is that Speaker 1 statements often still leave Speaker 2 statements with no response. While real-life conversations frequently display similar trends, a good dialogue dataset would ideally have a range of different interaction patterns.

And finally, because Reddit posts are from an online forum with others, our dialogues sometimes have phrases, diction, and tone that would be unusual to see in most spoken or written dialogue. This is crucial because our data is aimed at use by chat bot models which would then pick up those mannerisms, affecting its output quality.

6.1.3 Bias Propagation

A large concern when taking data from social media for use in training neural networks is bias. Unmoderated online content can often contain harmful language or stereotypes, and our model is currently unequipped to deal any sort of dangerous statement. Because of this, dialogues generated by our model should not yet be used for training neural networks without a proper understanding of the risks entailed.

6.1.4 Future Work and Difficulties

There are many possibilities for future work on this model. One good example is that of bias prevention: even just the use of a basic neural model to detect harmful language and stereotypes would help clean the output dialogues. As well, any of the methods listed in Section 2.2 would be good to investigate, however we did notice that most neural models we experimented with were less beneficial than the work we did around control flow, comment handling, and threading.

For this reason, future work on this model should be directed first into making better use of Reddit's features. For example, we never made use of the quoting capability or tagged questions/posts, and both of these would be good avenues to explore. Clever use of NLP models could of course overcome the limitations that we encountered with neural models, but the effort may be more significant than taking advantage of what already exists in Reddit.

However, a neural model capable of finding the best-suited middle sentence between two other sentences would be a great addition to this model. We use NSP scores to approximate a neural model that does this, but a trained model would surely perform much better. This would help smooth over transitions between speakers beyond what threading has managed to achieve.

An additional route of further work for this project could be a more developed evaluation, such as evaluating against databases of higher quality such as Switchboard or matching dialogues of similar topic to further remove bias from the grading.

As well, pre-filtering and post-filtering should be better examined. Pre-filtering could be done by taking posts and evaluated generated dialogues and training a model to recognize when a post will likely result in a low-quality conversation. Post-filtering is currently done by GRADE, set up by Daniil Huryn, but more post-filtering using different automated metrics could help moderate the quality of final dialogues included in generated datasets. GRADE is only currently used to remove low-quality dialogues, as we found that it didn't perform well at identifying high-quality conversations for our data.

The model could also be made faster by better management of comment deletion. Because we potentially have partial comments as possible threading responses as well as Speaker 2 utterances, we currently have to search through all possible comments with each sentence within the partial comment to properly prevent repetition. This could be better done given a system of hash maps for comments and threads.

6.2 Data Analysis

We quickly discovered that the quality of the data makes a big impact on the quality of the output dialogues, especially as generative neural models did not fit well into our approach. Because of this, different sub-reddits and their varying post styles can have strong effects on our assembled conversations.

We tended to notice that opinion/fact-dense sub-reddits lead to worse conversations. This is because our model, being context-apathetic, becomes more likely to sequence contradictory statements within a set of a speaker's utterances due to the sheer quantity of claims made in all of the content it is using to create the dialogue.

Additionally, sub-reddits that rely mainly on links, photos, or videos are entirely inaccessible to our model. Modifications to the model could be made to account for this, but as of right now we cannot handle such posts.

As well, on any sub-reddits where comments or posts have specific syntax and forms, our model's dialogues are impacted accordingly. A good example of this is in r/books there are often bi-weekly posts with the format: 'Simple Questions: February 19, 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?'. On these sorts of posts, our model creates poor dialogues, so there's a potential for big improvement given subreddit-specific pre-processing, which we did not employ.

For these reasons, when generating data from a sub-reddit using this model, the sub-reddit's post and comment style and typical content must be examined first to ensure the output will be of a reasonable quality.

6.3 Data Examples

To get a better sense of the type and quality of data we produced, what follows are analyses of a few dialogues we generated.

Dialogue one is a good example of the average dialogue found in our generated data. Transitions between statements are for the most part not awkward, though it becomes clear after several turns that Speaker 2 doesn't respond to Speaker 1 in any meaningful way. In this dialogue, threading did not kick in for any of the comments, and this is what causes the lack of interaction to stand out.

Dialogue two serves as an example of high-quality dialogue found in our data. Transitions are not awkward at all, and each speaker seems to build on each other's statements. Threading is used twice here by Speaker 1, both times are emphasized in the text for easier viewing.

Dialogue three is a conversation that exhibits the flaws of our final approach. Several Speaker 1 statements are uncomfortably long in comparison to Speaker 2 utterances and a comment is chosen for the first Speaker 2 utterance that responds to the post in its entirety. Beyond this, threading does not help smooth the transitions and there is not any interaction between speakers. Most dialogues that we remove in post-filtering look like this one, but we are not yet able to catch them all.

Dialogue One - Medium Quality

S1: "How do you come to terms with the fact that life isn't long enough to read all the books you want to read? The list of books I'm interested in reading probably would take two or three human lifetimes to get through (by that I mean a standard human lifetime involving other life goals like career, travel, raising kids, engaging in hobbies)."

S2: "It's not that life isn't long enough to read all the books you want to, it's that there are so many awesome books that you'll literally never run out."

S1: "The impossibility of reading everything I ever want to read due to my limited lifespan has triggered a kind of existential crisis in my enjoyment of reading."

S2: "Oftentimes when picking a book out I'm using a strange sense of intu-

ition. I feel like it always leads me to the books that are essential reads to aid in my growth/self actualization. Now this may very well be some new age, hippie dippy bs, but it makes me feel better about all of the great books that I'll never get around to."

s1: "I'm always trying to get to the next book, so I've been rushing my way through books without fully being absorbed in them."

s2: "My theory is to always have more books on my TBR shelf then I can possibly read in the next 5 years. That way I always have something to match my current mood. Due to the subjectivity of the experiance of reading a book, coming to terms with not reading them all is more about accepting the directionn of your path in life."

s1: "I guess I'm posting this to see if anyone here has shared this feeling and has any advice about it?"

s2: "I read a lot of books I really enjoy, even so the idea of having a TBR pile that extends more than a few months is a bit puzzling for me because I generally consume books far quicker than they are created. I currently have zero books on my TBR pile. I Honestly consider myself spoiled because I've read so many books that I enjoyed tremendously that its difficult for new books to measure up."

s1: "What?"

Dialogue Two - High Quality

s1: "What is the most disturbing book you've read? When I was 15 I was looking around the YA section of my local library looking for any paranormal books as it was my favorite genre at the time."

s2: "I was 13 or 14 and found a box of my dad's old horror paperbacks.....",

s1: "*Oof!* I came across Living Dead Girl by Elizabeth Scott and thinking it was a ghost story or something I picked it up and took it home."

s2: "Lord of the Flies...I was 15."

s1: "(I never read what the book was about bc I liked to be suprised.)"

s2: "But the most disturbing is Lemonade by Nina Pennacchi. I'm not sure if the author wrote it to be disturbing or if it was meant to be a romance, but.... yeah there's nothing romantic about it."

s1: "*What's it about?*"

Dialogue Three - Low Quality

s1: "Reading fiction books. Good morning, community."

s2: "The advice is ultimately really simple: read the books. Pay attention to them. Make an active effort to do so.",

s1: "The reason why I am writing this message is to ask you for advice about my situation below In my country, my class and other classes have been chosen to deep-read 10 fictional books in 3 months, write an analysis about them, and basically, after all of that go to a place where you debate/discuss with other nominated classes about the 10 books and decide which one of them was the best."

s2: "The 10 fiction books are already chosen by the teachers, and we're basically forced to read all of them regardless of whether we enjoy them or not, then write an analysis about each individual book."

s1: "I am not a guy that knows how this will work for me."

s2: "Do you have any limits on what books you can pick, or is any fiction acceptable? I say first find a book that you actually enjoy."

s1: "It's very common for me to read a book and then q̈uitït after reading the first 5 chapters of the book.. If I've somehow managed to finish reading a book, then the chances of me either not understanding or/and forgetting everything in a couple of days is very high with that in mind, I am planning

to make this like this: when reading the first book, I will write a summary of each chapter that I finish reading before going to the next chapter (if that made any sense) , and then after finishing the book, I will write an entire analysis of the entire book, and use the short summaries that I have written as a help-kit."

s2: "Basically, the books are already chosen and the entire class is forced to read them."

6.4 Analysis of Dialogues from Comparison Datasets

In Appendix B, one conversation from each of the target comparison datasets is listed. Where our data excels in comparison is in the naturalness of language we use.

In Topical-Chat's dialogue example, over half of the utterances begin with the word "Yes" or "Yeah", and facts are introduced at uncomfortable times in unnatural ways: "Did you know Bruce Lee was a cha cha dancer?" followed by "Yes he even won a hardcore cha cha championship in 1958". There is consistency and truth to the statements, but they are phrased and sequenced in such a way that they don't resemble real dialogue between humans.

In the MultiWOZ example, the dialogue is good but regimented. The speakers both use formal diction, so the dialogue wouldn't serve as a good example for a conversational chat bot. This aside, each utterance strongly responds to the previous utterance without any awkwardness. MultiWOZ is by far the most consistent dataset of the three we compared against.

The DailyDialog conversation showcases many of the same issues as the Topical-Chat dialogue, just at a lower frequency. Though the interaction percentage is high, statements such as, "If they are willing , we could ask them to go dancing with us. That is excellent exercise and fun , too ." cause some of the DailyDialog conversations

to score lower.

Our dialogue avoids many of the issues of DailyDialog and Topical-Chat because our phrasing was not generated by crowdsource workers. Posts and comments on Reddit are of a high enough caliber that even imperfect sequencing results, on average, in a conversation that reads as slightly more natural than those of DailyDialog and much more natural than Topical-Chat.

Chapter 7

Conclusion

Overall, our generated data scored very well in comparison to other dialogue datasets. Despite a few shortcomings due to the design of the model and the nature of the base data, our generated dialogues managed to outperform dialogues from several popular dialogue datasets of good quality. Our evaluation suggests that our dialogues could be used to successfully train a chat bot, or even be used solely to fine-tune it on a specific topic. Additionally, this data could be used for models being trained to rank dialogue options or for any number of other dialogue-related NLP tasks. We are currently looking into writing a conference paper for our work as well as providing an API for researchers to be able to use our model to generate data.

Perhaps equally as important is the fact that assembly dialogue models show great promise. Our model, though it has a complex control flow, does not involve complicated NLP techniques. Proper application of such techniques could greatly improve upon the results of our model, possibly bringing the quality of automatically generated dialogue to that of real-world human interactions.

Appendix A

Core Approach Pseudocode

```
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


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()
    Call getRandomPost() and store the data into a post_data variable
    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
    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

Comparison Dataset Example Dialogues

B.1 Topical-Chat

S1: do you like dance?

S2: Yes I do. Did you know Bruce Lee was a cha cha dancer?

S1: Yes he even won a hardcore cha cha championship in 1958

S2: Yeah. Did you know Tupac was a ballet dancer?

S1: Yes and he even was in the production of the nutcracker

S2: Yeah. Ballet dancer go through 4 pairs of shoes a week

S1: Yes that is a lot of shoes and also a lot of money

S2: Yeah true. Did you know babies are really good at dancing?

S1: Yes and they smile more when they hit the beat

S2: Yeah they are much smarter than we give them credit for

S1: True Did you know Jackson had a patent on a dancing device?

S2: Yes it helped him smooth out his dance moves

S1: Nice. Do you like Shakespeare?

S2: Yes I do. Do you know that he popularized many phrases

S1: Yes like good riddance, in my heart of hearts and such

S2: Yes and then he also invented names like Jessica, Olivia and Miranda

S1: Yes. And for his works you have to use old english for it to make sense

S2: Yes otherwise the rhymes and puns do not seem to work out

S1: Yes. He lived at the same time as Pocahontas too

S2: I wonder if they met how that would go from there

S1: Yeah interesting point. Nice chat

B.2 MultiWOZ

S1: I would like a taxi from Saint John's college to Pizza Hut Fen Ditton.

S2: What time do you want to leave and what time do you want to arrive by?

S1: I want to leave after 17:15.

S2: Booking completed! your taxi will be blue honda Contact number is 07218068540

S1: Thank you for all the help! I appreciate it.

S2: You are welcome. Is there anything else I can help you with today?

S1: No, I am all set. Have a nice day. Bye.

S2: you too! thank you

B.3 DailyDialog

S1: Say , Jim , how about going for a few beers after dinner ?

S2: You know that is tempting but is really not good for our fitness .

S1: What do you mean ? It will help us to relax .

S2: Do you really think so ? I don't . It will just make us fat and act silly . Re

S1: I guess you are right.But what shall we do ? I don't feel like sitting at home

S2: I suggest a walk over to the gym where we can play singsong and meet some of o

S1: That’s a good idea . I hear Mary and Sally often go there to play pingpong.Per

S1: Sounds great to me ! If they are willing , we could ask them to go dancing wit

S2: Good.Let ’ s go now .

S1: All right .

Appendix C

Example Dialogue from Each Approach

First Approach

S1: "Why do classes give out so much work? & I'm not just talking about moving to online, either"

S2: "I'm not even working rn because of my schoolwork amount"

S1: "Ever since my very first semester at college, my professors have been pilling on the reading homework and quizzes assignments back to back to back"

S2: "I can relate to this as the quizzes and questions are so heavy it's like the point is just do get them done not even use ur critical thinking skills"

Beam Search BERT NSP Approach

S1: "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)"

s2: "Yes, I have at least 3, 7-10 page readings I have to write about a week."

s1: "I've been reading anthropology articles all day."

s2: "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."

Final Approach - High Quality

s1: "What is the most disturbing book you've read? When I was 15 I was looking around the YA section of my local library looking for any paranormal books as it was my favorite genre at the time."

s2: "I was 13 or 14 and found a box of my dad's old horror paperbacks.....",

s1: "*Oof!* I came across Living Dead Girl by Elizabeth Scott and thinking it was a ghost story or something I picked it up and took it home."

s2: "Lord of the Flies...I was 15."

s1: "(I never read what the book was about bc I liked to be suprised.)"

s2: "But the most disturbing is Lemonade by Nina Pennacchi. I'm not sure if the author wrote it to be disturbing or if it was meant to be a romance, but.... yeah there's nothing romantic about it."

s1: "*What's it about?*"

Appendix D

Punctuation Excluded

```
.’’,!()?\n -;:~\"&$%+\"*/
```

Appendix E

Example Post and Comments

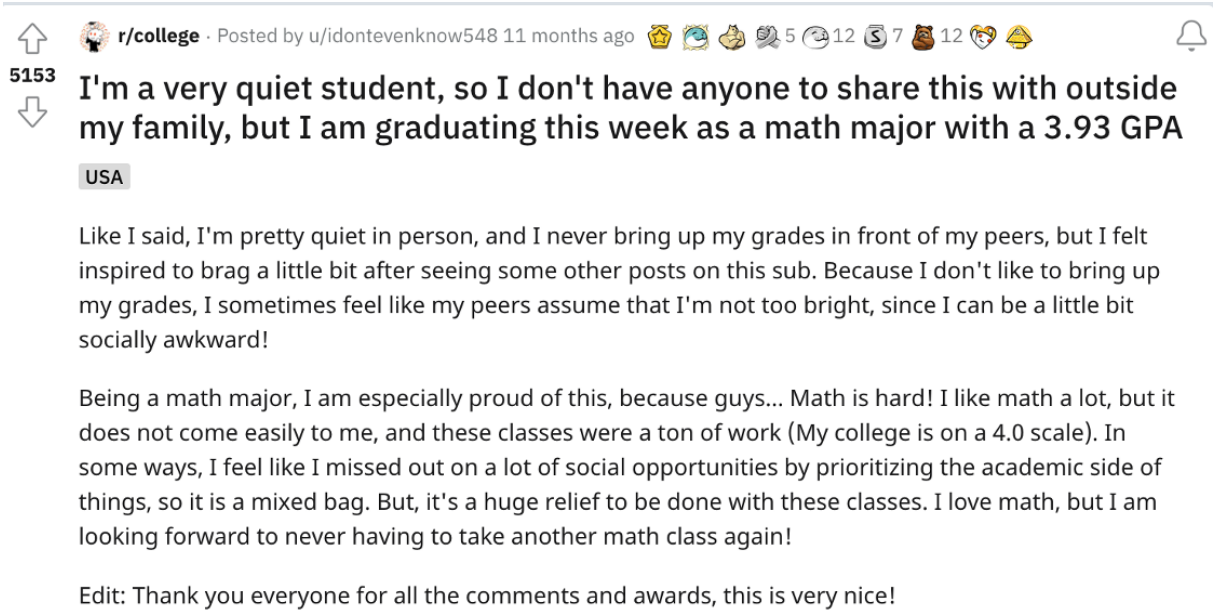

Figure E.1: Reddit Post View

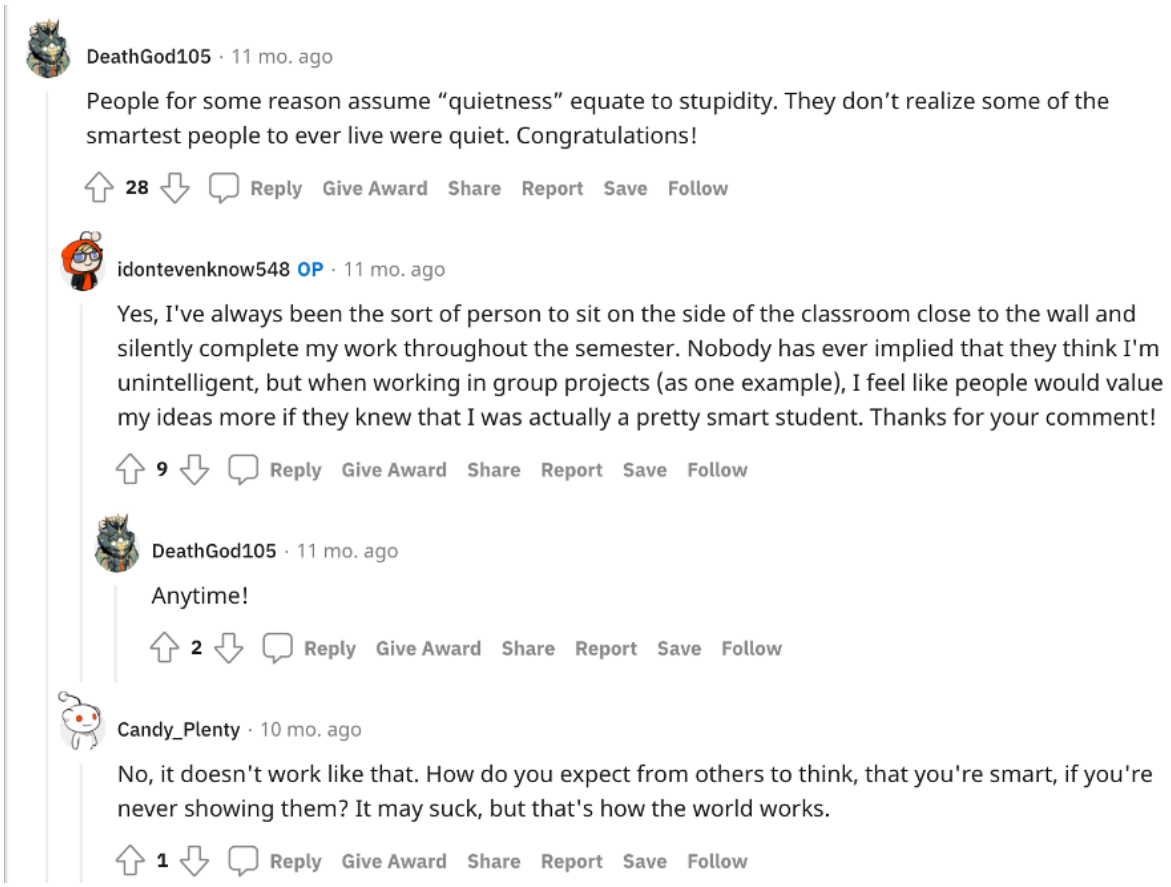

Figure E.2: Reddit Comments View

Appendix F

Amazon Turk Detailed Results

Even broken down into exact, coarse, and one-off agreement, we were not able to eliminate randomness from affecting our Turk data. Exact agreement was when both annotators agreed on the same label. Coarse agreement was based on grouping 1 and 2, 3 by itself, and then 4 and 5. One-off agreement is when their ratings were simply a difference of one away. The results are displayed in Tables F.1, F.2, and F.3. "More Natural" captured the largest percentage in exact agreement results, then dropped slightly in coarse agreement results, then fell more for One-off agreement results; this is because it was the most often guessed answer, which made it most likely to have overlap with the other annotator.

Table F.1: Exact Agreement Results (Total 122/400)

Score	Number	Percentage
1 - Significantly Less Natural	0	0%
2 - Less Natural	16	13.11%
3 - The Same	21	17.21%
4 - More Natural	80	65.57%
5 - Significantly More Natural	5	4.1%

As a last proof, we set up a smaller Turk task of 100 tasks in which we always set conversations from other databases first, and the results became much more

Table F.2: Coarse (and Exact) Agreement Results (Total 344)

Score	Number	Percentage
1 - Significantly Less Natural	1	0.29%
2 - Less Natural	33	9.59%
3 - The Same	42	12.21%
4 - More Natural	209	60.76%
5 - Significantly More Natural	59	17.15%

Table F.3: One-Off (And Exact) Agreement Results (Total 554)

Score	Number	Percentage
1 - Significantly Less Natural	1	0.18%
2 - Less Natural	63	11.37%
3 - The Same	147	26.53%
4 - More Natural	284	51.26%
5 - Significantly More Natural	59	10.85%

favorable toward other datasets, as shown in Table F.4. Note that these scores were flipped, 1 to 5, 2 to 4, as to showcase the results in the same set-up as previous tables, even though in the Turk sanity test we had workers rate the other datasets dialogues in relation to ours. All of these conversations had been scored in the opposite order in the previous turk task, yet over 90% showed completely different scores in this test.

Table F.4: Turk Sanity Test (Total 200)

Score	Number	Percentage
1 - Significantly Less Natural	11	5.05%
2 - Less Natural	102	51%
3 - The Same	52	26%
4 - More Natural	31	15.5%
5 - Significantly More Natural	4	2%

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] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241, 2018. URL `http://arxiv.org/abs/1811.01241`.

[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] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566, 2021. URL https://arxiv.org/abs/2107.07566.

[8] 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.

[9] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL http://arxiv.org/abs/1409.3215.

[11] 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.

[12] Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. A comprehensive assessment of dialog evaluation metrics. *CoRR*, abs/2106.03706, 2021. URL https://arxiv.org/abs/2106.03706.

[13] 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.

[14] 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.