

Diversity in Dialogue Generation

Katherine Stasaski



Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-48

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-48.html>

May 1, 2023

Copyright © 2023, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Diversity in Dialogue Generation

by

Katherine Stasaski

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Marti Hearst, Chair

Professor John DeNero

Professor David Bamman

Spring 2022

Diversity in Dialogue Generation

Copyright 2022
by
Katherine Stasaski

Abstract

Diversity in Dialogue Generation

by

Katherine Stasaski

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Marti Hearst, Chair

Neural conversational dialogue agents often produce uninteresting, broad responses, such as “Yes” or “I don’t know.” While these responses can be appropriate in a variety of contexts, if a model over-produces these typical responses, this leads to a dull conversation. This well-documented phenomenon is known as the diversity problem. This dissertation examines the diversity problem and proposes ways to improve dialogue agents in both the single- and multi-response setting.

In the single-response setting, the dialogue model is tasked with generating one utterance to continue a conversation. In this setting, a dialogue model’s diversity is measured by its ability to generate diverse responses to different conversations. I propose a data collection procedure aimed at increasing the diversity of a corpus, called Diversity-Informed Data Collection (DIDC). While prior work modifies decoding procedures to increase model diversity, DIDC addresses the diversity problem at the dataset level. DIDC uses dynamically computed corpus-level statistics to determine which conversational participants to collect more data from. DIDC produces significantly more diverse data than baseline data collection methods. Additionally, training dialogue models on a more diverse corpus results in more diverse responses. DIDC is generalizable and can be used with other corpus-level metrics.

The next two contributions consider the task of generating multiple responses for a single conversation. Diversity examined in this setting measures a model’s ability to generate multiple varied responses for the same input. First, I propose a novel metric which uses Natural Language Inference (NLI) to measure the semantic diversity of a set of model responses for a conversation. I evaluate this metric using an established framework and find strong evidence indicating NLI Diversity is correlated with semantic diversity. I show that the contradiction relation is more useful than the neutral relation for measuring this diversity. I additionally demonstrate how to iteratively improve the semantic diversity of a sampled set of model responses via a new generation procedure called Diversity Threshold Generation, which results in an increase in NLI Diversity compared to standard generation procedures.

Finally, I hypothesize that some conversations constrain the type of responses which are appropriate, therefore limiting the diversity one would expect in a set of responses. I explore the relationship between speech acts present in the input conversation and the diversity of a set of output responses. I propose the concept of Pragmatically Appropriate Diversity, the extent to which a conversation creates and constrains the creation of multiple diverse responses. Using a multi-response dataset, I find significant differences among NLI Diversity of responses for different speech act utterances. I use these findings to explore whether expert creative writers can predict the Pragmatically Appropriate Diversity from an input conversation, finding significant differences between the Pragmatically Appropriate Diversity among different speech acts. This contribution provides a framework to incorporate pragmatic conversational information into the evaluation of neural dialogue models.

*In loving memory of
Ron Griffith and Martha Stasaski.*

Contents

Contents	ii
List of Figures	iv
List of Tables	vi
1 Introduction	1
1.1 Measuring Diversity	3
1.2 Improving Diversity	4
1.3 Dialogue Diversity and Pragmatic Appropriateness	4
1.4 Dissertation Contributions	4
1.5 Statement of Prior Publication	5
2 More Diverse Dialogue Datasets via Diversity-Informed Data Collection	7
2.1 Introduction	7
2.2 Related Work	8
2.3 Diversity-Informed Data Collection	10
2.4 Evaluation	16
2.5 Discussion	20
2.6 Conclusion	21
3 Semantic Diversity in Dialogue with Natural Language Inference	23
3.1 Introduction	23
3.2 Related Work	25
3.3 NLI Diversity Metric	28
3.4 Evaluation of NLI Diversity	29
3.5 Diversity Threshold Generation	34
3.6 Evaluation of Diversity Threshold Generation Method	35
3.7 Discussion	42
3.8 Conclusion	43
4 Pragmatically Appropriate Diversity for Dialogue Evaluation	44
4.1 Introduction	44

4.2	Related Work	46
4.3	Speech Act Analysis	48
4.4	Study with Creative Writers	56
4.5	Discussion and Future Work	67
4.6	Conclusion	68
5	Conclusion	69
5.1	Contributions	69
5.2	Future work	70
	Bibliography	71

List of Figures

1.1	Hypothetical conversation which illustrates the diversity problem. Although the chatbot produces responses which are coherent, the responses are not engaging and do not further the conversation.	2
1.2	Illustration of Test Set and Multi-Response diversity evaluation conditions. While the Test Set diversity evaluation compares diversity for a model’s single response for each conversation, the Multi-Response Diversity condition measures the model’s ability to capture varied responses for each conversation.	2
1.3	Two conversations selected from DailyDialog++ (Sai et al., 2020; Y. Li et al., 2017) with different levels of Pragmatically Appropriate Diversity.	5
3.1	Illustration of NLI Diversity using human responses from DailyDialog++. Contradictions are weighted by 1, entailments by -1, and neutrals by 0, so the score is $(2 \times 1) + (3 \times 0) + (1 \times -1) = 1$	24
3.2	Spearman’s Correlation with 95% Confidence Intervals.	32
3.3	Histogram of Confidence NLI Diversity for high and low semantic diversity examples.	33
3.4	Histogram of number of samples required before reaching intended number of contradictions. Each bar color represents a different run of the experiment.	41
4.1	Two conversations from DailyDialog++ with sets of responses from writers (Section 4.4). When considering different ways the speaker could continue the conversation, the Pragmatically Appropriate Diversity is higher for Conversation A (Yes-No Question) than for Conversation B (Apology).	45
4.2	NLI Diversity comparison for responses categorized by most-recent speech act utterance (higher values indicate more diverse, ordered by diversity). Mean values are indicated by the white circle and corresponding text label. Box-and-whisker plots show the interquartile ranges.	52
4.3	Differences in NLI Diversity means between different speech acts. * indicates significance with $p < 0.05$. Darker blue rows indicate the Act2 speech act is more diverse than the Act1 speech act.	53

4.4	Sent-BERT comparison for responses categorized by most-recent speech act utterance (higher values indicate more diverse, ordered by diversity). Mean values are indicated by the white circle and corresponding text label. Box-and-whisker plots show the interquartile ranges.	54
4.5	Differences in Sent-BERT Diversity means between different speech acts. * indicates significance with $p < 0.05$. Darker blue rows indicate the Act2 speech act is more diverse than the Act1 speech act.	55
4.6	Participant progression through sections of the PA Diversity study.	56
4.7	Writing task from Qualtrics survey.	57
4.8	Drag and drop task from Qualtrics survey.	58
4.9	Likert rating task from Qualtrics survey.	59
4.10	Drag-and-drop results for Set 1 (above) and Set 2 (below).	65
4.11	Likert results for Set 1 (above) and Set 2 (below).	65
4.12	Averaged Likert ratings compared to NLI Diversity scores of DailyDialog++ responses. Because NLI Diversity is discrete, we added a random jitter for all points between -0.2 and 0.2.	66
4.13	Averaged Likert ratings compared to Sent-BERT Diversity scores.	67

List of Tables

2.1	Participant Population Selection conditions for Simulated Data Collection. The first three conditions are used in conjunction with Algorithm 1, while the last condition provides an upper-bound for diversity by utilizing a Corpus-Wide Oracle to determine the known most-diverse utterances.	13
2.2	Diversity metrics considered for data collection.	14
2.3	Results for diversity scores for each method of collecting corpora, by metric (Outlier, Entropy, and Mean IDF). Higher scores are better for all metrics. Also shown are the average number of participants (Avg. #Part) included out of a possible 810. * indicates statistical significance compared to the Random Population baseline ($p < 0.001$).	17
2.4	Results for downstream classification accuracy averaged over 5-fold cross-validation over 10 trials: higher is better. The task is classification of emotions from a set of 32 possible given the text of dialogue responses in <i>sub_c</i> . † and * indicate $p < 0.05$ and 0.001 respectively compared to Random Population.	18
2.5	Downstream model generation results; higher numbers are better for all metrics. †, §, and * indicate $p < 0.05$, 0.01, and 0.001 respectively. As Distinct-1 and Distinct-2 are summary statistics, we did not test significance.	20
3.1	Descriptions of diversity datasets from Tevet et al. (2021). Corresponding temperature parameter (higher is more diverse) or semantic and lexical diversity levels accompany each example.	30
3.2	Spearman’s ρ correlations between NLI Diversity metrics and the diversity parameter. Results above the double line are reproduced from Tevet et al. (2021). Both the best automatic metric and human performance for each dataset are in boldface.	31
3.3	Spearman’s ρ correlation between NLI Diversity metrics (MNLI) and human judgments. Negative values indicate higher # Entailment is <i>negatively</i> correlated with diversity.	34
3.4	Diversity results of using Diversity Threshold Generation (with a <i>div_{thresh}</i> of 10 # Contradictions for NLI, 0.98 for distinct-n, and -0.164 for Sent-BERT). Num. sampled has a maximum value of 20; DG is the DialogGPT model; BB is BlenderBot.	36
3.5	Average utterance overlap from starting to ending set of responses using Diversity Threshold Generation on multiple models, datasets, and diversity metrics.	37

3.6	Starting and ending responses with corresponding NLI Diversity for DialoGPT, randomly sampled from the DailyDialog++ dataset. SD = Starting NLI Diversity; ED = Ending NLI Diversity.	40
3.7	Results comparing starting and ending sets of responses from Diversity Threshold Generation to sets of human responses using two relevancy metrics, BERTScore and BLEU score.	42
4.1	Speech acts from DailyDialog and SWBD along with a sample conversational utterance from DailyDialog++. DailyDialog++ speech act definitions are quoted from Amanova et al. (2016). For brevity, we only include one turn for each speech act instead of the entire conversation.	49
4.2	Diversity for multi-response sets responding to different speech acts along with the number of conversations. For both metrics, higher values indicate higher level of diversity.	50
4.3	Average NLI and Sent-BERT Diversity among sets of participant-produced responses for speech act categories. Speech acts are ordered within each set based on our hypothesized diversity level, from most to least diverse.	61
4.4	Writer responses created in response to each speech act. To directly compare how a writer’s responses might change based on speech act, we selected one writer’s responses to Set 1 speech acts (Yes-No Question, Wh-Question, Apology, and Thanking) and a separate writer’s responses to Set 2 speech acts (Open Question, Opinion, Non-Opinion, and Closing).	64

Acknowledgments

This dissertation would not have been possible without the help and support of many individuals. I'd like to first thank my wonderful advisor Marti Hearst, for being instrumental in shaping my research abilities the past six years. Thank you for encouraging me to explore a wide variety of problems, for challenging and pushing me to improve my ideas, and for helping me become a better researcher, teacher, and mentor.

I'd also like to thank the other members of my dissertation committee, David Bamman and John DeNero, as well as Bjorn Hartmann, who served on my qualifying exam committee. I appreciate the feedback and thoughts on this work that I've gotten from each of you. I'd like to thank my collaborators Grace Hui Yang, who helped shape the content in Chapter 2, and Orianna DeMasi, who first inspired me to think about the diversity problem in dialogue. Thank you also to Vikram Ramanarayanan, Mark Hopkins, and Ronan Le Bras, for mentoring me in research outside of Berkeley.

All of my labmates in Marti's group have been incredibly helpful in providing diverse perspectives and pushing my work further. To Nate Weinman, thank you for all of the friendship, support, and encouragement you've given me over the years. Thank you to Philippe Laban, Andrew Head, and Dongyeop Kang, for both friendship and providing valuable feedback on this work. Thanks to Chase Stokes, Peitong Duan, and the other Hearst Lab group members for the feedback and support of this work. I'd also like to thank the fantastic undergraduate students that I've had the pleasure of mentoring: Julian Park, Kimberly Kao, Kiran Girish, Kevin Lu, Avni Prasad, Emily Xiao, Tony Tu, and Manav Rathod. Thank you for pushing me to become a better mentor and being a pleasure to work with.

My Berkeley NLP labmates have provided incredibly thought-provoking conversations about NLP over the years, especially David Gaddy, Samee Ibraheem, Daniel Fried, Cathy Chen, Nikita Kitaev, Mitchell Stern, Nick Altieri, Jacob Andreas, Nick Tomlin, and Lucy Li. Thank you to Armando Fox, An Ju, Dan Garcia, Nick Weaver, Michael Ball, Pamela Fox, Hezheng Yin, Kristin Stephens-Martinez, Zephyr Barkan, Vron Vance, and the entire ACELab community for the interesting conversations on educational research and improving CS teaching. Thanks to Molly Nicholas, Sarah Serman, Eldon Schoop, Jeremy Warner, James Smith, and all of my colleagues in the BiD lab, who have provided valuable insight in human-computer interaction and a wonderful community.

Thanks to my collaborators from the George Washington University, who first exposed me to NLP and Education research, Rahul Simha and Jennifer Pizzurro, as well as my other undergraduate mentors, Bhagi Narahari, Poorvi Vora, Tim Wood, Gabe Parmer, and Michael Clarkson. Thank you for encouraging me to go to graduate school.

I am thankful for my funding sources: a National Science Foundation (NSF) Graduate Research Fellowship, a UC Berkeley Chancellor's Fellowship, an AWS Machine Learning Research Award, an AI2 Key Scientific Challenges grant, and an NVIDIA Corporation GPU grant.

I'd finally like to thank my family and friends for providing a support network outside of graduate school. My incredible partner, Aakash Thakkar, has been by my side supporting me every step of this journey. My friends and community, including Rachel Chen, Terrence Lewis, and Leena Thakkar, among many others have been incredibly supportive as well. Finally, my wonderful parents Mike and Renea, grandparents Helen, Ron, and Martha, brother Denny and sister-in-law Shelby have been my rock. Thank you for supporting me throughout this journey.

Chapter 1

Introduction

Conversational chatbots aim to engage and entertain humans via social conversations. In contrast to goal-oriented dialogue systems, which have an explicit goal such as booking a plane ticket, chatbots generate conversational responses without an overarching task. Chatbots instead aim to entertain and engage users in conversational chat. For example, one social chatbot, incorporated into the Amazon Alexa software, was able to keep users engaged in conversation for an average of 7.33 minutes (Yu et al., 2019).

In addition to entertainment benefits, chatbots have the potential to be an engaging educational exercise. While this dissertation does not explore the use of social chatbots for education, prior educational work with chatbots has found that interacting with a chatbot has benefits for student engagement and motivation in language-learning activities (Fryer et al., 2020; Fryer et al., 2019; Fryer et al., 2017; Fryer et al., 2006; Ki et al., 2019). In particular, Fryer et al. (2006) found that 85% of undergraduate students learning English preferred practicing their skills with a chatbot compared to a peer and 74% found a social chatbot entertaining. Fryer et al. (2017) found students are initially as interested in a language learning task when conversing with a chatbot as with a human, although this interest declines over time. Chatbots provide the opportunity for students to practice language skills without fear of being judged by their peers and with a conversational partner that is always available.

Many early chatbots, such as such as ELIZA (Weizenbaum, 1966), were based on a series of pattern- and keyword-matching rules (Ki et al., 2019). However, more recently, neural networks have greatly improved the performance and generalizability of conversational chatbots (Y. Zhang et al., 2020; Roller et al., 2021). Neural dialogue agents are trained or fine-tuned on large corpora of conversational data and have the potential to generalize from training dataset to unseen data. However, in practice, past work has shown these models over-produce the more frequent responses represented in this data (J. Li et al., 2016a). Responses such as “Yes,” “No,” or “I don’t know” are over-produced by neural chatbots. While these typical responses are applicable in a wide variety of contexts, conversing with a model which over-generates these responses is not engaging for the user.

Consider the hypothetical conversation in Figure 1.1. While the chatbot does not say anything ungrammatical or which is incoherent given the conversation, all produced responses

Human: I just got promoted at work!
Neural Chatbot: Okay.
Human: How should I celebrate?
Neural Chatbot: I don't know.

Figure 1.1: Hypothetical conversation which illustrates the diversity problem. Although the chatbot produces responses which are coherent, the responses are not engaging and do not further the conversation.

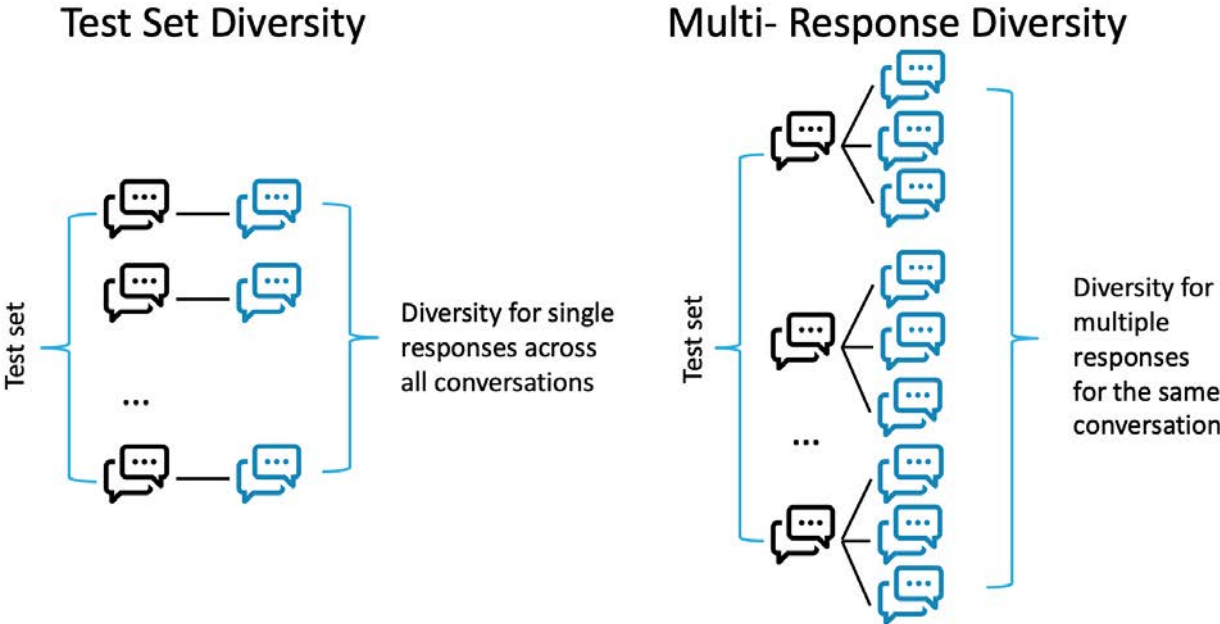


Figure 1.2: Illustration of Test Set and Multi-Response diversity evaluation conditions. While the Test Set diversity evaluation compares diversity for a model's single response for each conversation, the Multi-Response Diversity condition measures the model's ability to capture varied responses for each conversation.

are dull. The chatbot does not engage the user and does not further the conversation. Thus, it follows that when evaluating conversational chatbots, there should be a measure which captures a model's ability to produce interesting, varied, and appropriate responses.

1.1 Measuring Diversity

The most common condition used to evaluate a chatbot model’s diversity involves the chatbot generating a single response for each conversation in a held-out test set, which I refer to as *Test Set Diversity*. Figure 1.2 (left) provides an illustration of this condition. A diversity metric is computed over the model’s responses, producing a single diversity score. Formalized, this calculation is:

$$diversity_{test_set} = div_{metric}(\forall_{x \in test_set} response_{model}(x))$$

There have been several metrics proposed to automatically measure diversity in the Test Set condition. J. Li et al. (2016a) propose measuring the proportion of unique n-grams generated normalized by the total n-grams generated. A similar metric measure the number of distinct responses over the test set (Cao et al., 2017). BLEU score can also be calculated among model responses generated from the test set (Zhu et al., 2018).

The Test Set diversity setting can provide indication if the model is over-generating common responses across multiple conversations in the test set. However, we might want to capture, instead, if the model is able to generate *multiple* distinct responses for the same input conversation. This allows us to determine if the model can capture different perspectives and provide varied options for the next response in a conversation.

Diversity metrics in the Multi-Response setting typically take in multiple responses for the same conversation and produce a diversity score, resulting in a score for each conversation in the test set. Figure 1.2 (right) provides an illustration of this condition. These scores can be averaged over all test set conversations to report a single number. Formalized for a model which generates i responses from each input conversation in the test set:

$$diversity_{multi_response} = \frac{1}{len(test_set)} \sum_{x \in test_set} div_{metric}(response_{0_{model}}(x), \dots, response_{i_{model}}(x))$$

Test Set diversity metrics can be repurposed in the Multi-Response setting by treating each set of responses to a conversation as its own evaluation set (Tevet et al., 2021). Other metrics in the Multi-Response diversity setting consider the similarity of embeddings among sets of model responses, assuming similarity to be inversely correlated with diversity (Tevet et al., 2021).

Depending on the intended application of a dialogue model, one of these conditions may be more important than the other. This dissertation makes contributions in both the Test Set and Multi-Response diversity conditions.

1.2 Improving Diversity

Several strategies have been proposed aimed at increasing the diversity of neural model responses. For example, decoding methods such as nucleus sampling (Holtzman et al., 2020), where words in an utterance are sampled from a dynamically sized set of words which together contain at least p probability mass, can improve diversity. Backtranslation, where an utterance is first translated to a different language and then back to the original language, has been used to improve fluency and diversity of generated text (Dhole et al., 2020). Paraphrase generation, where models are directly trained to paraphrase a sentence, has also been used to increase diversity (Jia et al., 2020). Other approaches have added constraints or incentives to the text a model can produce at decode time (Baheti et al., 2018; J. Li et al., 2016b; J. Li et al., 2017; Shao et al., 2017) or use a temperature parameter to sharpen the output vocabulary distribution (Cao et al., 2017).

In contrast to this past work, this dissertation explores a new way of improving model diversity by creating a more diverse *dataset* (Chapter 2). I also propose a novel decoding task aimed at iteratively improving the diversity of a set of responses (Chapter 3).

1.3 Dialogue Diversity and Pragmatic Appropriateness

The past sections have established that is important for chatbots to be able to generate diverse responses. However, maximal diversity should not always be desired, particularly in the Multi-Response setting. Some conversations contain underlying speech acts, which constrain the pragmatic appropriateness of next responses (Levinson, 1983). Consider the two conversations in Figure 1.3, selected from a multi-response dialogue dataset. The thanking speech act present in Conversation B constrains the types of responses which are pragmatically appropriate, restricting the diversity we would expect to see from a set of responses continuing the conversation.

While diversity of chatbot responses is important, it is also important to be aware that diversity may be constrained by a conversation’s underlying speech acts. This dissertation analyzes the interactions between speech acts and chatbot diversity in the Multi-Response setting (Chapter 4).

1.4 Dissertation Contributions

In Chapter 2, I propose a new technique to improve model diversity in the Test Set diversity setting. In contrast to past work which alters the chatbot model or decoding process used, I propose the task of collecting a more diverse *dataset* to train chatbot models on. Based on the hypothesis that some crowdworkers create more diverse data than others, I propose a novel dataset collection algorithm to measure each crowdworker’s contribution to a corpus.

Conversation A

Speaker 1: Have you adapted to the lifestyle in China?

Speaker 2: Yeah, almost. But there is one thing.

Speaker 1: What?

Conversation B

Speaker 1: Excuse me! How can I get to the post office please?

Speaker 2: Well, turn left at the first corner after the crossroad. It's near the corner. You can't miss it.

Speaker 1: Thank you very much.

Figure 1.3: Two conversations selected from DailyDialog++ (Sai et al., 2020; Y. Li et al., 2017) with different levels of Pragmatically Appropriate Diversity.

Using this new dataset collection process results in significantly more diverse datasets than a traditional collection process. This also translates to downstream model performance; models trained on the more diverse datasets also produce more diverse responses to conversations.

In Chapter 3, I switch from the Test Set setting to the Multi-Response setting. I propose a novel diversity metric (NLI Diversity) based on the predictions of a Natural Language Inference model. Using an established diversity metric evaluation framework, I find high correlation between NLI Diversity and human judgments of semantic diversity. I also propose a new decoding procedure, which iteratively improves the diversity of a set of model responses with minimal loss in relevancy.

In Chapter 4, I explore the relationship between dialogue diversity and the underlying speech acts of the conversation being responded to. I formulate the notion of Pragmatically Appropriate Diversity, the extent a conversation inspires and constrains the creation of multiple responses. I find significant differences for the diversity of human-generated sets of multiple responses for different speech act utterances. I also propose a new task where writer participants rate the Pragmatically Appropriate Diversity of different conversations, finding judgments correspond to speech act hypotheses.

1.5 Statement of Prior Publication

Chapter 2 has been previously published at the Association for Computational Linguistics conference in 2020 (Stasaski et al., 2020). My coauthors Marti A. Hearst and Grace Hui Yang, who helped conceptualize this work and prepare the publication, have provided consent

for this material to be reproduced in this dissertation. Chapter 3 has been accepted for publication at the North American Association for Computational Linguistics conference in 2022. My coauthor Marti A. Hearst helped formulate this work and has given permission for this material to be reproduced in this dissertation. Finally, Marti A. Hearst has been instrumental in the conceptualization and presentation of the ideas presented Chapter 4.

Chapter 2

More Diverse Dialogue Datasets via Diversity-Informed Data Collection

2.1 Introduction

Automated generation of conversational dialogue using modern neural architectures has made notable advances. However, these models are known to have a drawback of often producing uninteresting, predictable responses; this is known as the diversity problem. We introduce a new strategy to address this problem, called Diversity-Informed Data Collection.

Unlike prior approaches, which modify model architectures to solve the problem, this method uses dynamically computed corpus-level statistics to determine which conversational participants to collect data from.

Diversity-Informed Data Collection produces significantly more diverse data than baseline data collection methods, and better results on two downstream tasks: emotion classification and dialogue generation. This method is generalizable and can be used with other corpus-level metrics.

It is well-documented that neural dialogue models struggle with generating engaging, relevant responses (Jiwei Li et al., 2016) and often produce banal responses such as “Yeah.” While this may be an appropriate response to a chitchat conversation, to keep a human participant engaged, diversity of responses is important. Diverse models vary the language used and the content referenced, and the generated utterances differ from the most typical conversation responses some proportion of the time. A model which only generates “Yeah,” “No,” and “I don’t know” is not diverse and is not be engaging to converse with.

Past work has improved model diversity with innovation on model architectures and decoding strategies (Jiwei Li et al., 2016; Baheti et al., 2018; J. Li et al., 2017; Shao et al., 2017; Cao et al., 2017; Serban et al., 2017; Zhao et al., 2017). We build upon this work to propose a novel method to collect and determine more diverse data to train these models with. Our method can be used in conjunction with existing generation-specific model innovations.

Some prior work on data collection processes has prioritized diversity. For instance,

Rashkin et al. (2019) prompts crowdworkers to choose an underused emotion class to generate dialogue. This work encourages coverage of emotion classes, but does not consider the likelihood that some crowdworkers are better at producing certain types of data than others.

This chapter introduces Diversity-Informed Data Collection (DIDC), a new strategy for creating a dataset of conversational utterances via selecting which participants’ data to include in the collection. The strategy progressively builds up a more diverse sub-corpus from an existing larger collection. The main idea is to grow the sub-corpus by adding conversations sequentially and to assess the contribution of a new participant’s utterances to the diversity of the entire sub-corpus. This strategy is also applicable to on-the-fly collection of new datasets via crowdworking or similar methods. We implement DIDC with three diversity metrics: Outlier, Entropy, and Mean-IDF.

Diversity-Informed Data Collection also provides a new method for finding an upper bound on a current corpus’s diversity via a Corpus-Wide Oracle which has access to information about which utterances are most diverse across the corpus.

Prior work has not used corpus-level statistics to enhance the diversity of the collected data. Instead, when collecting data with crowdworkers, researchers have sought more diverse responses by altering the task (Kang et al., 2018) or by altering the stimulus (Larson et al., 2019). Prior work that trains neural dialogue models has not made use of subsets of existing datasets that exhibit properties of diversity.

Our experiments show this strategy yields significantly more diverse data than baseline collection processes. It also yields better, more diverse model output on two downstream tasks. Additionally, this method can be implemented for other metrics which are defined relative to the corpus.

2.2 Related Work

Past work in neural dialogue generation investigates how to improve diversity in conversational responses. Additionally, past work in crowdsourcing data collection has explored optimizing crowdsourcing data collection processes.

Diverse Neural Dialogue Generation

Improving model diversity is an important goal in dialogue generation (Jiwei Li et al., 2016), with several related works proposing architecture and training improvements to increase diversity.

Decoding methods to increase model diversity include Jiwei Li et al. (2016) which proposes maximizing mutual information between the source sentence and response rather than maximizing likelihood. Other approaches have focused on beam search and incentivizing diverse beams, by adding similarity constraints at decoding (Baheti et al., 2018), penalizing items on the beam that are similar and reranking resulting items (J. Li et al., 2016b), or penalizing words which have already been generated in a current beam (J. Li et al., 2017).

Shao et al. (2017) uses attention over already-generated words at decode time and beam reranking. Adding a temperature parameter to sharpen the decoder’s distribution has also been studied (Cao et al., 2017).

Neural architecture improvements have also been explored, such as conditioning on a latent variable at decode time (Serban et al., 2017; Zhao et al., 2017) or a multi-headed attention mechanism which aims to capture different parts of the context (Tao et al., 2018). Y. Zhang et al. (2018) explore the use of Generative Adversarial Networks to incentivize diversity. These more diverse models and decoding methods can be used in conjunction with Diversity-Informed Data Collection, since it attempts to improve the data that neural models are trained on in an earlier part of the model pipeline.

Crowdsourcing

Related work in crowdsourcing has approached the optimization problem of how to assign crowdworkers to different tasks.

Crowdworker Task Assignment

Basu Roy et al. (2015) formulates the problem of matching crowdworkers to tasks depending on skill levels for a set of concepts, pay rates, and HIT acceptance ratio. Follow-up work extends to collaborative crowdwork, where crowdworkers need to work together (Rahman et al., 2015). Assadi et al. (2015) pursue a similar task assignment setup.

Additional work has attempted to automatically evaluate crowdworker quality of task performance and use the results to assign crowdworkers to new tasks on-the-fly (Fan et al., 2015). Further investigations have explored more adaptive assignment of tasks in real-time based on the likelihood that a participant will continually complete tasks (Kobren et al., 2015). Relatedly, Kumai et al. (2018) design a task allocation to minimize the stress of workers and maximize the resulting quality in terms of balanced skill performance.

Label Distribution Prediction

An additional area related to our work is crowdworker label distribution prediction. T. Liu et al. (2019) has a crowdworking labeling task and trains models to predict the 50-label crowdworker distribution from 5-10 labels. Yang et al. (2018) aim to predict diversity in crowdworker answers to questions about an image to determine how many crowdworker responses are required to capture this diversity.

Dynamic Crowdworking Tasks

Lin et al. (2018) tackle the task of employing crowdworkers to generate or label minority class examples to feed an active-learning model. They deploy a multi-armed bandit to choose crowdworking tasks based on how cheaply a minority-class example can be generated using the technique. Our approach, by contrast, adapts a distributional constraint across the

entire collection. Zhou et al. (2018) explores the related task of changing crowdworker team instruction prompts.

Diverse Crowdfunding

Data collection approaches to incentivize diverse crowdworker output have also been studied. For instance, in EmpatheticDialogues (Rashkin et al., 2019) crowdworkers are conditioned to generate a response and an emotion (such as “afraid” or “proud”) associated with it. If workers do not generate text with certain emotions, they are prompted to select only from the underused labels. This is an example of trying to get better class coverage, but does not compare crowdworker output to the entire corpus of collected responses.

Past work has also examined how the particular crowdworking task affects the diversity of crowdworker output. Kang et al. (2018) compare two crowdsourcing tasks for use in a downstream goal-oriented dialogue system and examine resulting data diversity. While Kang et al. (2018) focus on choosing a *task* which produces diverse utterances, our work focuses on choosing a *participant population* which produces diverse data compared to data which has already been collected.

Building on Kang et al. (2018), and perhaps most similar to our work is Larson et al. (2019), which tackles the problem of detecting outlier paraphrases generated by crowdworkers. To obtain multiple ways of expressing similar intent (such as opening a bank account), crowdworkers are asked to paraphrase sentences. After a round of paraphrase collection, the most diverse (the outlier) paraphrases are identified and placed back onto the crowdsourcing platform for another round of data collection.

Our method is similarly aimed at increasing diversity of collected data. However, our method adapts the participant population for a set of tasks, which can be used in addition to an approach like Larson et al. (2019) which adapts the stimulus the population works on.

2.3 Diversity-Informed Data Collection

We propose a method, **Diversity-Informed Data Collection**, which progressively builds up a corpus, and while doing so, identifies which conversation participants produce more diverse utterances compared to the rest of the in-progress corpus. More formally, our task is to progressively build a sub-corpus, sub_c , of a given size from a larger, pre-collected corpus, c , where utterances are tied to IDs of specific participants.

Our approach is aimed at building a diverse sub-corpus sub_c . Our approach chooses which *population* of participants to collect data from for a given round. This population changes dynamically depending on calculated participant’s diversity scores.

When utilizing a human-created, pre-existing corpus, we assume responses of the dataset are well-formed and of acceptable quality. With this assumption, we can maximize diversity scores without worrying that quality will be sacrificed for this diversity. However, when

using this approach to collect data on-the-fly, additional quality controls may be necessary to ensure diverse data does not come at the cost of quality.

We assess two experimental conditions: Simulated Data Collection and Corpus-Wide Oracle Upper-Bound. Simulated Data Collection is set up to mimic crowdsourcing data collection processes leveraging a large pre-collected corpus, while Corpus-Wide Oracle Upper-Bound gathers an maximally diverse sub-corpus of utterances.

Corpus

For all experiments, we utilize the pre-collected EmpatheticDialogues corpus (Rashkin et al., 2019). We experiment with this corpus because it has crowdworker IDs associated with each utterance, which allows us to experiment with varying the participant population. Future work should conduct further experimentation to examine this approach’s adaptability to other chitchat and goal-oriented datasets.

The corpus has a large number of utterances (100,000) over 25,000 conversations. Each conversation is centered around a situation (such as getting a promotion at work) and is associated with one of 32 emotions, such as anger, excitement, or guilt. Each conversation takes place between two crowdworkers and is an average of 4.3 turns. There are 810 unique crowdworkers in this dataset, each completing an average of 132 utterances each across an average of 61 conversations.

Our task is to select sub_c of size 10,000 from the larger EmpatheticDialogues corpus, c . We choose 10,000 as it is a sufficient number of utterances to train downstream models but still a small proportion (10%) of the original dataset, allowing examination of differences between sub-corpora. Implementation utilizes Cornell Convokit (Chang et al., 2019).

Simulated Data Collection

We simulate real-time crowdsourcing using a large, pre-collected corpus, c . This allows for running multiple trials, each time selecting sub_c and examining significance of different diversity metrics and participant selection conditions.

We simulate collecting data on-the-fly using an artificially-constructed environment (formally described in Algorithm 1), which completes multiple rounds of data collection until the progressively built sub-corpus $size(sub_c)$ is the desired size. The procedure assumes a fixed number of conversation participants in each round to gather data from (set to 10 for our experiments). We collect 2 conversations from each participant, chosen to allow the algorithm to recover from a participant with low diversity utterances while not judging a participant on just one conversation.

Given a participant’s conversation, the diversity of an utterance in that conversation is stated in Equation 1:

$$div_{utt} = ComputeDiversity(utt, sub_c) \tag{2.1}$$

Algorithm 1: Data collection simulation environment. *ComputeDiversity* depends on the diversity metric (Table 2.2), and *EvalParticipants* depends on the participant selection approach (Table 2.1).

```

1 function GatherData(Corpus c)
2   subc =  $\epsilon$ 
3   subCorpusSize = 10,000
4   numConvosToCollect = 2
5   population = []
6   numParticipants = 10
7   while size(subc) < subCorpusSize do
8     while size(population) < numParticipants do
9       p = Sample from c.Participants
10      population.append(p)
11      c.Participants.remove(p)
12    end
13    participantDiversities = []
14    for Participant p in population do
15      divp = 0
16      numUtts = 0
17      for i in numConvosToCollect do
18        convo = sample from p.Convos
19        for utt in convo do
20          divp += ComputeDiversity(utt, subc)
21          numUtts += 1
22          subc.append(utt)
23        end
24        p.Convos.remove(convo)
25      end
26      divp /= numUtts
27      participantDiversities.append(divp)
28    end
    // Which participants kept for next round based on diversity scores.
29    toKeep = EvalParticipants(participantDiversities)
    // Which participants still have data.
30    remaining = p in population where len(p.convos) ≥ numConvosToCollect
31    population = (toKeep ∩ remaining)
32 end

```

where ComputeDiversity depends on the diversity metric examined. We obtain a diversity score for each participant p 's set of utterances ($utts_p$) by averaging these diversity values:

$$div_p = \frac{1}{size(utts_p)} \sum_{utt \in utts_p} div_{utt} \quad (2.2)$$

At the end of each round of data collection, utt_p is added to sub_c for each participant. Additionally, the algorithm determines which subset of the participant population is retained for the next round based on a Participant Population Selection strategy.

Our algorithm is greedy, since the order participants are added to the simulation and the order in which conversations are sampled both affect the participant's likelihood to be retained for an additional round. However, crowdworker data collection itself is usually a greedy approach, with crowdworkers being assigned to tasks in the order they arrive and being allowed to complete many tasks until the dataset has been collected.

Condition	Description
Diverse Population	Calculates each participant's average relative diversity for current data collection round. We retain the 70% most-diverse participants of the current round.
Above Mean Population	Calculates each participant's average relative diversity for current data collection round. Retains the participants whose diversity scores fall above the sub-corpus's mean diversity.
Random Population	Retains a random 70% of participants.
Corpus-Wide Oracle	Uses a Corpus-Wide Oracle which ranks utterances' diversities in relation to the large dataset, c . Selects the most diverse utterances from these values independent of conversations.

Table 2.1: Participant Population Selection conditions for Simulated Data Collection. The first three conditions are used in conjunction with Algorithm 1, while the last condition provides an upper-bound for diversity by utilizing a Corpus-Wide Oracle to determine the known most-diverse utterances.

Participant Population Selection

We experiment with three conditions to determine which sub-set of *current participants* (participants which were involved in the most recent round of data collection) should be retained for the next round of data collection, summarized in Table 2.1.

Metric	Description
Outlier	Euclidean distance between utterance embedding and average embedding for all utterances in the sub-corpus (Larson et al., 2019)
Entropy	Entropy of utterance under a trigram language model trained on sub-corpus.
Mean IDF	Mean IDF value (Baeza-Yates et al., 1999) for words in utterance compared to the rest of the corpus.

Table 2.2: Diversity metrics considered for data collection.

Diverse Population: After collecting conversations from current participants, we choose to retain the most-diverse 70% of participants.

Above Mean Population: Any participant whose diversity average falls above the mean diversity average of sub_c is retained in the pool of participants.

Random Population: We compare to a special random baseline, where at each iteration we retain a random 70% of the participant population, to directly compare to the 70% of crowdworkers retained in Diverse Population. We structure Random Population to collect data from roughly the same number of participants as Diverse Population, to examine differences between the resulting sub_c due to the the *selection* of which participants to retain for another round of data collection.

Diversity Metrics

We experiment with three diversity metrics (Outlier, Entropy, and Mean IDF), summarized in Table 2.2. For all metrics, a new utterance utt is compared to the sub-corpus sub_c .

The same utterance can have different diversity values depending on the utterances in sub_c . When augmenting pre-collected data, this allows for the collection of new utterances which are *relatively* diverse.

Outlier: The embedding-based Outlier metric was proposed by Larson et al. (2019). Each utterance is encoded using a Universal Sentence Encoder (USE), which creates a sentence embedding by averaging word embeddings and passing the representation through a feedforward neural network, originally trained in a multi-task setting with supervised and unsupervised NLP tasks (Cer et al., 2018).

An embedding of an utterance is created via: $E_{utt} = USE(utt)$. A *mean corpus vector* is

computed by averaging all of sub_c 's utterance's vectors:

$$E_{sub_c} = \frac{1}{size(sub_c)} \sum_{u \in sub_c} USE(u) \quad (2.3)$$

The diversity metric is the Euclidean distance between each new utterance and the mean corpus vector, or:

$$\sqrt{\sum_i (E_{u_i} - E_{sub_{c_i}})^2} \quad (2.4)$$

where i is a dimension in Embedding E .

Utterances which are farther from the mean corpus vector are given a higher diversity score. For Simulated Data Collection, the mean corpus vector shifts as data is collected. Therefore, depending on which utterances are already added in the sub-corpus, outlier values will change for a given utterance.

Entropy: The Entropy score is determined by a non-neural trigram language model with smoothing for unseen words. The diversity score is given by:

$$-\frac{1}{|x \in Trigram(utt)|} \sum_{x \in Trigram(utt)} p(x) \log p(x) \quad (2.5)$$

The language model is only trained on utterances in the sub-corpus.

Mean IDF: This metric calculates the mean IDF value for each word in the utterance (Baeza-Yates et al., 1999). IDF is calculated by treating each utterance in the corpus as a document. For a given utterance utt_p and sub-corpus sub_c , Mean IDF is calculated via:

$$\frac{1}{|utt_p|} \sum_{w \in utt_p} \log \left(\frac{|\{sub_c\}|}{|\{utt|w \in utt\}|} \right) \quad (2.6)$$

where $\{sub_c\}$ is the set of all utterances in the sub_c . The IDF of a word w in utt is the number of utterances in sub_c divided by the number of utterances containing w on a log scale.

In addition to evaluating the robustness of our approaches, multiple diversity metrics are chosen with different conceptual types of diversity in mind. Outlier uses Universal Sentence Encoder embeddings which capture content (Cer et al., 2018). Entropy considers the probability of short phrases and can capture word combination diversity. Mean IDF considers the rarity of words being used for vocabulary diversity. Depending on the downstream application for a dialogue agent, the utility of these diversity measures may vary.

Corpus-Wide Oracle Upper Bound

To provide an Upper Bound for the diversity of a sub-corpus sub_c , we create a Corpus-Wide Oracle which knows the value of each utterance's diversity compared to the entire corpus c .

For each $utt \in c$, we compute diversity according to the methods in Table 2.2, where $sub_c = c$. For example, for Outlier, the mean corpus vector is

$$\frac{1}{size(c)} \sum_{x \in c} USE(x) \quad (2.7)$$

which captures utterances from the entire corpus c . We calculate a Corpus-Wide Oracle diversity score, div_{oracle} , for each utterance in c for each diversity metric.

The Corpus-Wide Oracle is used to construct sub_c of any size consisting of the most diverse utterances. This sub-corpus can be used to compare against other collection methods, such as those in Simulated Data Collection, or as a way to enhance an existing collection by selecting out the most diverse utterances.

After the Corpus-Wide Oracle ranks each utterance by diversity, we select the utterances with the top 10,000 diversity values to form sub_c . This serves as a use-case for collecting the maximally-diverse corpus for a given diversity metric.

However, the Corpus-Wide Oracle might not be the *best* 10,000 utterances to collect for a sub-corpus. The Corpus-Wide Oracle selects the utterances with the most diversity compared to the whole corpus, but this might be too much diversity without enough context since the Simulated Data Collection methods add entire conversations (not utterances in isolation) to sub_c .

2.4 Evaluation

We evaluate the collected corpora both in terms of how diverse each sub-corpus is as well as performance on two downstream tasks: conversation emotion classification and dialogue generation.

Overall Diversity

The first evaluation aims to answer the question of if our methods produce more diverse sub-corpora than the Random Population baseline. We examine the hypothesis that using a collection method with knowledge of diversity will result in sub_c that is significantly more diverse. For each data collection method, we compare the diversity of the sub-corpus to Random Population. Because diversity values are relative to sub_c , diversity of sub_c is measured via div_{oracle} values.

Table 2.3 shows the resulting div_{oracle} values for datasets collected using our methods. Each value is the average of 100 trials, in which each trial collects a 10,000 utterance sub-corpus, sub_c .

Significance results for all experiments use a two-sided t-test compared to the Random Population baseline. Both Diverse Population and Above Mean Population produce datasets which contain statistically significantly ($p < 0.001$) more diverse data compared to the Random Population baseline. The Corpus-Wide Oracle method produces the most diverse

	Condition	Mean Score	Avg. #Part
Outlier	Random Population	0.974	257.4
	Diverse Population	0.979*	262.1
	Above Mean Population	0.978*	516.9
	Corpus-Wide Oracle	1.035*	539.0
Entropy	Random Population	-5.350	257.2
	Diverse Population	-5.320*	259.1
	Above Mean Population	- 5.294*	359.1
	Corpus-Wide Oracle	- 4.261*	481.0
Mean IDF	Random Population	5.455	256.2
	Diverse Population	5.659*	257.7
	Above Mean Population	5.613*	357.5
	Corpus-Wide Oracle	7.783*	546.0

Table 2.3: Results for diversity scores for each method of collecting corpora, by metric (Outlier, Entropy, and Mean IDF). Higher scores are better for all metrics. Also shown are the average number of participants (Avg. #Part) included out of a possible 810. * indicates statistical significance compared to the Random Population baseline ($p < 0.001$).

results overall, as expected as it is a collection of the top 10,000 most diverse utterances. Running Diversity-Informed Data Collection to collect datasets of size 5,000 produced similarly significant differences.

We also examine the average number of participants out of the 810 total in c that are included for each method. Note in Table 2.3 the difference in Average Number of Participants from Random Population and Diverse Population to Above Mean Population and Corpus-Wide Oracle. Even though Above Mean Population is more diverse than Diverse Population for Entropy, it comes at the cost of more participants. Across all three diversity metrics, Above Mean Population requires about 100–200 additional participants than Diverse Population and Random Population. In an online setting where the cost to train new crowdworkers is high, the tradeoff between number of participants and diversity of content may be worth considering.

Classification

To examine the quality of the resulting sub_c 's, we turn to downstream task evaluation. We first examine the task of classifying a conversation's emotions from utterance text. Following Larson et al. (2019)'s justification, we would expect more diverse sub_c to result in higher classification accuracies, because more diverse responses should cover more variation in how people express emotions in conversation.

	Condition	SVM	Fast-Text
Outlier	Random Population	0.224	0.050
	Diverse Population	0.234*	0.052
	Above Mean Population	0.229	0.077*
	Corpus-Wide Oracle	0.100*	0.057*
Entropy	Random Population	0.218	0.052
	Diverse Population	0.212†	0.049
	Above Mean Population	0.254*	0.065*
	Corpus-Wide Oracle	0.134*	0.102*
Mean IDF	Random Population	0.220	0.052
	Diverse Population	0.236*	0.052
	Above Mean Population	0.257*	0.064*
	Corpus-Wide Oracle	0.131*	0.065*

Table 2.4: Results for downstream classification accuracy averaged over 5-fold cross-validation over 10 trials: higher is better. The task is classification of emotions from a set of 32 possible given the text of dialogue responses in *subc*. † and * indicate $p < 0.05$ and 0.001 respectively compared to Random Population.

Classification Method

We follow the methodology of Larson et al. (2019) who propose evaluating the diversity of goal-oriented intent paraphrases. For their use case, classification models predict the intents from the paraphrase. For our case, each conversation in the EmpatheticDialogues corpus is associated with an emotion, such as anger or guilt. There are 32 such emotions throughout the corpus. The classification task is to predict which of the 32 emotions is expressed from a given utterance. Following Larson et al. (2019), we use two classification models:

- Bag-of-Words SVM
- FastText classifier

Bag-of-Words SVM is an SVM using TF-IDF word features for prediction. The FastText classifier uses a neural classification model on top of fastText sentence embeddings (Joulin et al., 2017). The sub-corpora we collect using the different methods serve as the datasets to train these classification models.

Classification Results

Classification task results are summarized in Table 2.4. Reported scores are averaged 5-fold cross-validation and averaged over 10 runs of datasets collected from each method.

While most conditions show Diverse Population significantly outperforms Random Population, it performs worse than Random Population with Entropy SVM and Entropy FastText

and performs the same in Mean IDF FastText. Above Mean Population, on the other hand, outperforms the Random Population baseline on all conditions. This could potentially be due to the larger number of participants included in Above Mean Population. Surprisingly, Corpus-Wide Oracle does not perform the best in each category. We conjecture that too many diverse responses do not allow a classification model to learn common characteristics.

Generation

Because the ultimate goal of collecting more diverse dialogue data is generating more diverse text, we evaluate diversity of neural text generation models trained on resulting corpora.

Generation Method

Our task is to generate the next utterance in a dialogue, where the data collection processes collect utterances for sub_c . To train generation models, the input is the most recent parent utterance for each utt in sub_c , and utt is the target sentence to generate. When utt is the starting utterance in a conversation, the input is the situation associated with the conversation (such as planning a vacation).

We train Sequence-to-Sequence models (Sutskever et al., 2014) with a 2-layer bidirectional encoder, hidden size 500, word vector size 64, Adam optimizer (Kingma et al., 2014), learning rate 0.001, trained for 3000 steps with batch size 32. Models are implemented using OpenNMT (Klein et al., 2017). We opt to use a standard model as it has fewer parameters to learn from smaller sub-corpora. We use the same parameter settings for all trained models.

Generation Results

Generation task results are summarized in Table 2.5. We report on both mean and median length of model responses. Distinct-1 and Distinct-2 measure the proportion of unigrams and bigrams respectively in the set of model responses which are unique (Jiwei Li et al., 2016). We also report diversity of the generated responses calculated by the metrics used in sub_c collection (see Table 2.2).

Our method results in models which produce more diverse output compared to baseline Random Population data collection. Interestingly, Diverse Population and Above Mean Population split the win on producing more diverse outputs. Corpus-Wide Oracle diversity results are sometimes lower and overall shorter in length than other methods; a potential reason is this condition only samples utterances, not conversations.

Responses from the model trained on each sub_c are evaluated with all 3 diversity metrics, to examine potential interactions. Collecting sub_c with Entropy results in higher Mean IDF (and vice versa) compared to Random Population. Collecting sub_c with Outlier results in slightly lower Mean IDF (and vice versa) for Diverse Population and Above Mean Population compared to Random Population. There is not a consistent signal between Outlier and Entropy. Future work can further examine the relationships among these diversity metrics.

	Condition	Mean Length	Median Length	D-1	D-2	Outlier	Entropy	Mean IDF
Outlier	Random Population	7.6	7	0.114	0.296	0.981	-3.088	5.504
	Diverse Population	9.7	7	0.110	0.279	0.989*	-3.354*	5.297§
	Above Mean Population	8.1	7	0.063	0.169	0.960*	-3.083	5.067*
	Corpus-Wide Oracle	3.8	4	0.204	0.448	1.042*	-2.968*	6.789*
Entropy	Random Population	8.8	8	0.101	0.265	0.981	-3.281	5.263
	Diverse Population	7.7	7	0.122	0.317	0.978	-3.197§	5.411†
	Above Mean Population	6.6	6	0.092	0.226	0.982	-3.057*	5.474*
	Corpus-Wide Oracle	4.9	5	0.112	0.316	0.985§	-2.935*	5.781*
Mean IDF	Random Population	6.1	6	0.120	0.294	0.988	-3.036	5.526
	Diverse Population	6.7	6	0.131	0.322	0.986	-2.955§	5.797§
	Above Mean Population	7.2	7	0.071	0.187	0.976*	-2.937*	5.655
	Corpus-Wide Oracle	3.4	3	0.214	0.449	1.008*	-2.421*	8.327*

Table 2.5: Downstream model generation results; higher numbers are better for all metrics. †, §, and * indicate $p < 0.05$, 0.01, and 0.001 respectively. As Distinct-1 and Distinct-2 are summary statistics, we did not test significance.

2.5 Discussion

Diversity Considerations: Compared to a random baseline, Diversity-Informed Data Collection results in more diverse data than Random Population, which is shown to be more effective on downstream tasks. Future work can explore the effect of simultaneously optimizing multiple desirable measurements of diversity.

However, we acknowledge that maximum diversity might not be what is desired and does not always result in the best downstream task performance, as indicated by the low

Corpus-Wide Oracle downstream task performance. While we have not examined the tradeoff between diversity and quality, this can be explored in future work.

Generalizability: Diversity-Informed Data Collection is generalizable to metrics other than diversity. Concretely, DIDC should be used when a desired metric (1) can compare one sample (or set of samples) to the in-progress dataset and (2) has variation among the participant population.

Additionally, Diversity-Informed Data Collection can be applied to areas outside of dialogue data collection. For instance, DIDC could apply to collecting data with different emotions or sentiment. Another extension is to a specialized application domain, such as collecting dialogues for educational tutoring purposes, where our method could be used to collect more data from students who generate text consistent with certain types of misconceptions.

Crowdworking Deployment: We evaluated on simulated crowdworking data by leveraging an existing corpus. This choice stems from the desire to test multiple runs of methods in a controlled environment, to reliably determine significance, and to work with data with an assumed level of quality. That said, our approach can be applied to real crowdworking tasks. Data can be gathered from several participants in parallel, where crowdworkers are added and offered new tasks or assigned qualifications based on their diversity.

If our method is deployed in paid crowdworking tasks, Diverse Population might be more cost-effective. In this particular investigation, we find Diverse Population requires 100-200 fewer participants than Above Mean Population to create a dataset. Due to the time required to train new participants, there is a tradeoff between training a new worker and collecting more data from current participants.

Caution should be taken in using this method on-the-fly without a quality check. Standard quality control methods (e.g., crowdworker qualifications, manual examination, crowdworker verification) should be deployed for from-scratch data collection.

Crowdworker Fairness: Another important consideration for a live deployment is the crowdworker’s perspective of fairness. Because some crowdworkers are retained for more data collection than others, communicating this possibility to crowdworkers is essential (Brawley et al., 2016). Crowdworking best practices involve disclosing which quality metrics are being used to workers to set clear expectations (Bederson et al., 2011). Additionally, combining our method with a method which alters the task crowdworkers complete (Kang et al., 2018) as opposed to restricting the crowdworking population could be a way to balance fairness with crowdworkers. Different task and population combinations could allow for all crowdworkers to participate in more tasks.

2.6 Conclusion

We propose a method, Diversity-Informed Data Collection, which leverages this to produce more diverse datasets than the standard approach, and which performs better on downstream tasks. We define diversity of an utterance compared to the other utterances in a corpus. This

allows for measurement of the impact of adding each utterance to the corpus. Working under the same assumption that a subset of participants produce diverse data compared to the corpus, our method can be extended to other diversity measures and can be modified to work with other corpus-level metrics.

Chapter 3

Semantic Diversity in Dialogue with Natural Language Inference

3.1 Introduction

While Chapter 2 proposed a way of improving a chatbot model’s diversity in the Test Set condition, this chapter instead considers measuring and improving a model’s diversity in the Multi-Response condition. In the Test Set condition, some prior work uses human evaluation to measure model response diversity according to engagingness, specificity, or interestingness (J. Li et al., 2016a; See et al., 2019; Ghandeharioun et al., 2019); however, several automated metrics have also been proposed to measure diversity of model responses. Some metrics measure *lexical* diversity, typically via n-gram overlap (J. Li et al., 2016a) or computing the BLEU score (Zhu et al., 2018) among model responses generated from the test set. Other past work attempts to measure *semantic* diversity via repurposing sentence similarity metrics (Tevet et al., 2021; T. Zhang et al., 2020; Cer et al., 2017).

We propose a new metric aimed at measuring semantic diversity by leveraging a Natural Language Inference (NLI) model to score a set of multiple dialogue model responses for a single conversation, as illustrated in Figure 3.1. NLI is a three-way classification task to determine whether one sentence entails, contradicts, or is neutral toward a second sentence. We hypothesize that a diverse set of responses for a conversation captures contradictory ways one could respond, which can be measured by the NLI model. We aggregate the contradiction, neutral, and entailment predictions among pairs of responses from the set and combine the predictions into a new diversity metric, called *NLI Diversity*.

We additionally explore two modifications of NLI Diversity. First, because the neutral prediction may be indicative of diversity, we propose Neutral NLI Diversity, where neutral predictions are weighted the same as contradiction predictions. Second, since our Baseline NLI Diversity method does not take into account the *confidence* of the model’s prediction, we propose Confidence NLI Diversity, which aggregates the probability mass of the model’s predicted class instead of aggregating the number of predictions for each class.

The Original Conversation

Speaker 1: I hated the hand-me-downs to wear when I was a kid.

Speaker 2: That's understandable. Every kid likes new clothes.

A Set of Next Responses

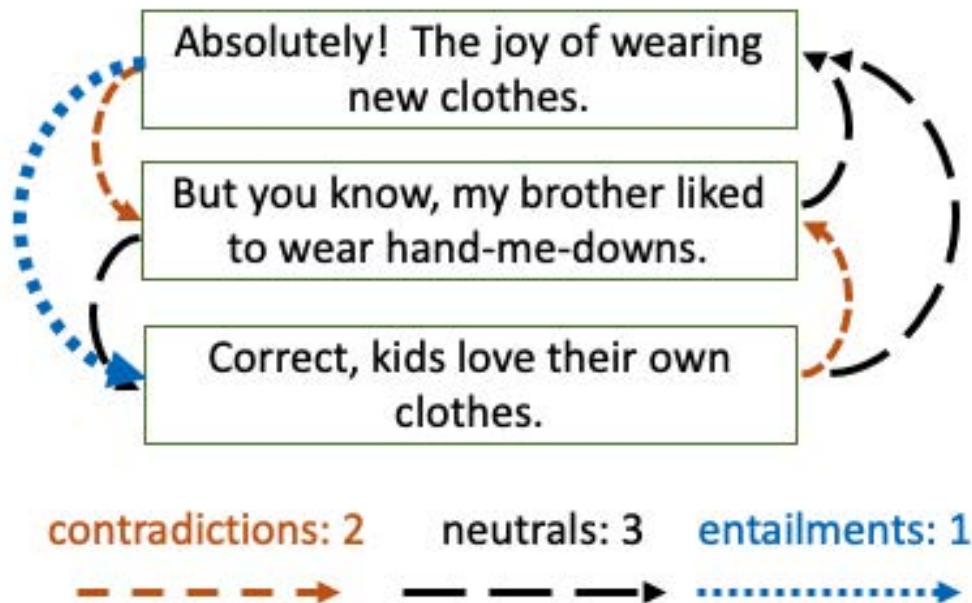


Figure 3.1: Illustration of NLI Diversity using human responses from DailyDialog++. Contradictions are weighted by 1, entailments by -1, and neutrals by 0, so the score is $(2 \times 1) + (3 \times 0) + (1 \times -1) = 1$.

We assess NLI Diversity using Tevet et al. (2021)’s diversity metric evaluation framework, finding that NLI Diversity is highly correlated both with human judgments of diversity and with the *diversity parameter*, a gold standard diversity value used to generate the set of responses. Confidence NLI Diversity achieves state-of-the-art performance in terms of correlation with semantic diversity. Also, through an ablation study, we find positive, neutral, and negative correlations between human judgments and the number of contradiction, neutral, and entailment predictions, respectively.

We next explore the use of a dialogue model to generate a set of candidate responses with a minimum target level of semantic diversity, such as 10 Contradictions. Our new generation procedure, *Diversity Threshold Generation*, iteratively improves a set of model responses until this intended threshold is reached. If a set of sampled responses does not meet the intended threshold, the lowest-scoring response is thrown out and a new response is sampled until the diversity threshold is reached. We show this procedure results in a more diverse set of responses than the original sampled set, often with only a few resampled responses. Results of automated analysis shows relevancy is maintained from initial to final sets of responses.

In summary, this chapter’s contributions are:

- A novel diversity metric, NLI Diversity, evaluated using Tevet et al. (2021)’s framework, that measures semantic diversity and interrogates the relationship between Contradiction and Neutral predictions and diversity,
- Confidence NLI Diversity, a diversity metric which obtains state-of-the-art performance on semantic diversity,
- A new dialogue generation procedure, Diversity Threshold Generation, which continues sampling responses until an intended diversity threshold, defined using NLI Diversity, is reached,
- Experimental results indicating dialogue models are able to generate diverse responses using Diversity Threshold Generation with minimal loss in relevancy.

3.2 Related Work

Past work has explored lexical and semantic diversity metrics as well as ways of evaluating these metrics. We also draw from work in NLI and generating diverse sets of hypotheses.

Measuring Model Response Diversity

Traditionally, a model’s diversity has been measured in terms of its predictions over the test set (J. Li et al., 2016a), which we call *Test Set Diversity*. In this setup, the model predicts one response for each conversation in the test set (containing n conversations), resulting in n predictions. The diversity measure is computed over these n predictions, resulting in a score over the entire test set.

The notion of diversity we investigate, however, measures the model’s ability to generate a *set* of responses for a single conversation (X. Zhang et al., 2019; Tevet et al., 2021),

which we call *Multi-Response Diversity*. Instead of generating one response for each of the conversations in the test set, we evaluate a model’s ability to generate m responses for each of the n conversations.

As shown by Tevet et al. (2021), metrics which have been proposed in the *Test Set Diversity* setting can still be applied in the *Multi-Response Diversity* setting, however, by treating each set of m responses as its own “test set” and averaging over the n total sets.

Diversity Metrics

Lexical diversity metrics measure differences in word choice, as opposed to diversity of content. J. Li et al. (2016a) propose *distinct-n*, which measures the number of unique n-grams generated divided by the total number of n-grams generated in the *Test Set Diversity* setting. Some past work has applied this metric to the *Multi-Response Diversity* setting (Tevet et al., 2021). Cao et al. (2017) propose examining the percent of unique *responses* over the test set. Other past work has proposed using BLEU score over a set of model responses in the *Test Set Diversity* setting (Zhu et al., 2018).

Semantic diversity metrics, on the other hand, compare diversity of the content present in each response. Many of these measures are adapted from semantic similarity scores, since lower similarity can indicate higher diversity (Tevet et al., 2021). BERTScore measures the similarity of BERT embeddings for each token in two sentences (T. Zhang et al., 2020). BertSTS assigns a score based on the semantic similarity of two sentences (Tevet et al., 2021). The Sent-BERT metric computes cosine similarity between BERT sentence embeddings (Reimers et al., 2019). Larson et al. (2019) propose identifying diverse paraphrases by identifying embedding outliers.

Other past work has used human evaluation to measure a model’s diversity. J. Li et al. (2016a) ask humans to choose the better of two responses based on specificity to the past conversation. See et al. (2019) ask humans to rank dialogue responses on a variety of factors, including interestingness and inquisitiveness. Tevet et al. (2021) compare participants’ ability to judge diversity of a set of responses in two ways: (i) by ranking one response as more diverse than a second response and (ii) by judging the diversity of a single response on a Likert scale, finding that participants were equally able to judge diversity in both conditions. They also find that human judges are better at distinguishing semantic diversity than lexical diversity.

Other past work has incorporated diversity metrics into the dialogue dataset creation pipeline. In Chapter 2, we propose a method which measures the diversity of a crowdworker’s contributions compared to a corpus, using that information to determine when to stop collecting data from the worker. This results in a more diverse dataset.

Evaluation of Diversity Metrics

Tevet et al. (2021) propose a framework to examine the reliability of diversity metrics. They propose the notion of a *diversity parameter*, which is used to generate a set of model responses,

e.g., the p -value in nucleus sampling, which specifies the vocabulary probability distribution cutoff used to restrict sampling to the most-likely words whose combined likelihood $\geq p$. If p is higher, the set of responses should have higher diversity, and vice-versa. This *diversity parameter* is treated as a gold standard for a set of responses' diversity. Diversity metrics assign scores in the *Multi-Response Diversity* condition and are evaluated in terms of correlation to the diversity parameter. They further propose two datasets to evaluate diversity metrics: one which includes model responses and contains varying levels of lexical diversity and one which is human-created and maintains high lexical diversity to allow focused evaluation of semantic diversity.

Natural Language Inference

Natural Language Inference is a task aimed at predicting whether one sentence contradicts, entails, or is neutral towards a second sentence. Models for NLI are typically trained using one of two datasets: Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) or Multi-Genre NLI (MNLI) (Williams et al., 2018). More recent datasets include FEVER (Thorne et al., 2018; Nie et al., 2019), adapted from a fact-checking dataset, and ANLI (Nie et al., 2020), collected in an adversarial human-in-the-loop procedure. With the rise of transformer architectures, models have achieved high performance on NLI tasks (Y. Liu et al., 2019).

In a dialogue setting, NLI has been used to improve *consistency* between a persona and model responses over the course of a conversation by integrating an NLI-based reward into a reinforcement learning training procedure (Song et al., 2020).

To our knowledge, however, NLI has not been used to measure the diversity of model responses in either the *Test Set Diversity* or the *Multi-Response Diversity* setting.

Generating Diverse Sets of Hypotheses

While work has only recently begun to explore the task of generating multiple *dialogue* responses to a conversation (X. Zhang et al., 2019; Tevet et al., 2021), past work has explored generating diverse sets of hypotheses in some other application areas. Carbonell et al. (1998) explored using Maximal Mutual Relevance to reduce redundancy without sacrificing relevancy in document selection for summarization. Batra et al. (2012) proposed a greedy iterative algorithm to generate diverse, probable hypotheses for multiple vision tasks. Most related to our work is Gimpel et al. (2013), which applied Batra et al. (2012)'s approach to machine translation, generating a *set* of translations instead of a single translation. In contrast to Gimpel et al. (2013), by holding the sampling procedure constant throughout the iterative process, our method can explore the extent to which diversity can be increased without altering standard decoding practices.

3.3 NLI Diversity Metric

We propose three diversity metrics in the *Multi-Response Diversity* setting which leverage the predictions of an NLI model. Two metrics (Baseline and Neutral) aggregate the NLI model’s class predictions and one metric (Confidence) aggregates the weight of these predictions.

Baseline NLI Diversity

We propose a new metric, called *Baseline NLI Diversity*, which uses an NLI model’s predictions to measure diversity. More formally, for a given conversation, c , and a dialogue generation model M , a set of utterances u_1, \dots, u_n is produced by the model. Each pair of utterances is compared in both directions using an NLI model, $NLI(u_1, u_2), NLI(u_2, u_1), \dots, NLI(u_n, u_{n-1})$.

The NLI model predicts a distribution over the three potential classes: contradiction, neutral, and entailment. We take the argmax over these classes, resulting in a list of NLI predictions, $NLI_{preds}(NLI(u_1, u_2), \dots, NLI(u_{n-1}, u_n))$ of size $n(n-1)$. To produce an overall diversity score for $NLI_{preds}(u_1, \dots, u_n)$, we assign each of these classes a value representing their diversity, denoted $NLI_{score}(NLI_{preds}(u_1, \dots, u_n))$.

We hypothesize that larger numbers of entailment predictions found in a set of model-generated utterances is indicative of a lack of diversity; similarly, larger number of contradiction predictions is indicative of a larger amount of diversity. Because we want a higher value of NLI_{score} to indicate higher diversity, we assign values as:

$$NLI_{score} = \begin{cases} 1 & \text{if contradiction} \\ 0 & \text{if neutral} \\ -1 & \text{if entailment} \end{cases}$$

The sum of the NLI_{score} values for the set of utterances results in the final NLI Diversity score, formally defined as:

$$\text{Baseline NLI Diversity} = \sum_{u_i, u_j \in u_1, \dots, u_n} NLI_{score}(NLI_{pred}(NLI(u_i, u_j)))$$

While the Baseline NLI Diversity metric aggregates all classes, we also investigate the separate number of entailment, contradiction, and neutral predictions in NLI_{preds} , denoted # Entailment, # Contradiction, and # Neutral, respectively.

Neutral NLI Diversity

Our primary hypothesis is that contradictions indicate diversity and entailments indicate lack of diversity. Because it is unclear what the role of neutrals might be, we explore a version of

NLI Diversity which weights neutral and contradiction predictions as equally diverse. This metric is the same as Baseline NLI Diversity except the NLI_{score} used to assign values is:

$$NLI_{score_neutral} = \begin{cases} 1 & \text{if contradiction} \\ 1 & \text{if neutral} \\ -1 & \text{if entailment} \end{cases}$$

Confidence NLI Diversity

Because the prior two NLI Diversity metrics do not incorporate the confidence of the NLI model’s class predictions, we explore an additional metric which incorporates this value. Letting $conf_{class}(u_1, u_2)$ represent the model’s probability mass assigned to the predicted NLI class after *softmax*, the function is defined as: $NLI_{score_confidence} =$

$$\begin{cases} 1 \times conf_{con}(u_1, u_2) & \text{if contradiction} \\ 0 & \text{if neutral} \\ -1 \times conf_{ent}(u_1, u_2) & \text{if entailment} \end{cases}$$

Intuitively, instead of assigning a 1 value for a contradiction prediction, this metric assigns the probability of the contradiction class. Likewise, instead of a -1 for an entailment prediction, this metric assigns the negative probability mass of the entailment class.

3.4 Evaluation of NLI Diversity

We evaluate NLI Diversity by computing the correlation between the metric and both human labels and *diversity parameter* labels. Below we first describe the models and data and then present the results of the evaluation.

Models

We explore two NLI models: a Roberta-large model (Y. Liu et al., 2019) fine-tuned on the Multi-Genre NLI (MNLI) Corpus (Williams et al., 2018)¹ and a Roberta-large model fine-tuned on a combination of MNLI, SNLI, FEVER, and ANLI², both containing 300M parameters. We refer to these models as *NLI Diversity – MNLI* and *NLI Diversity – Combined*, respectively. We do not employ additional fine-tuning of these models.

Data

There are two different English datasets released to evaluate diversity metrics in Tevet et al. (2021): *conTest* and *decTest*, described in Table 3.1. The *conTest* dataset is human-created

¹<https://huggingface.co/roberta-large-mnli>

²https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

decTest	Mixed Lexical Diversity; Mixed Semantic Diversity; Model Generated
Examples:	
temp 0.28	“I think he is the most awesome guy ever” “He is the most awesome guy ever”
temp 0.55	“The unemployment rate is lower than what it is” “No but it does make it more likely to be higher than what it is”
conTest	High Lexical Diversity; Mixed Semantic Diversity; Human Generated
Examples:	
high lexical and low semantic	“Sorry, but I don’t agree.” “I think you are wrong about that.”
high lexical and high semantic	“Dont be so judgemental, try to see things her way.” “You are right that is insane.”

Table 3.1: Descriptions of diversity datasets from Tevet et al. (2021). Corresponding temperature parameter (higher is more diverse) or semantic and lexical diversity levels accompany each example.

and captures *content*, or *semantic*, diversity independent of *lexical* diversity. Low-diversity examples in this dataset have high lexical diversity but low semantic diversity. This dataset was created by asking crowdworkers to generate sets of utterances with either low or high *semantic* diversity using varied language, in order to keep a high level of *lexical* diversity constant across both conditions.

The *decTest* dataset includes model-generated responses, with diversity controlled by a decoding parameter, such as a temperature parameter. The dataset can include duplicate responses, and does not attempt to mediate lexical diversity; therefore, low-diversity examples in this dataset may reflect *low lexical* as well as *low semantic* diversity.

While the original dataset includes multiple generation tasks, we evaluate on the dialogue task, *respGen*, which is drawn from Reddit conversations (Hashimoto et al., 2019)³. There are 200 conversations for each of *conTest* and *decTest* for the *respGen* task, with multiple responses for each conversation (5 for *conTest*, 10 for *decTest*).

³In the data released from Tevet et al. (2021), these files are called *con_test_200_with_hds_resp_gen.csv* and *dec_test_200_with_hds_resp_gen.csv* for *conTest* and *decTest*, respectively.

	decTest	conTest
Metric	ρ	ρ
Human Performance (absHDS)	0.81	0.63
distinct-n	0.89	0.34
cos-sim	0.89	0.33
BERT-STS	0.81	0.46
Sent-BERT	0.80	0.59
BERTScore	0.87	0.49
Baseline NLI Diversity – MNL	0.58	0.59
Baseline NLI Diversity – Combined	0.39	0.59
Neutral NLI Diversity	0.72	0.24
Confidence NLI Diversity	0.44	0.62

Table 3.2: Spearman’s ρ correlations between NLI Diversity metrics and the diversity parameter. Results above the double line are reproduced from Tevet et al. (2021). Both the best automatic metric and human performance for each dataset are in boldface.

Diversity Parameter Correlation

The *diversity parameter* from Tevet et al. (2021) represents either a parameter directly used to generate responses via a dialogue model, such as p in nucleus sampling, or a binary value indicating whether crowdworkers were instructed to generate a high- or low-diversity set of responses. A measure which is able to capture diversity will be positively correlated with this diversity parameter.

Table 3.2 shows Spearman’s correlations between NLI Diversity and the diversity parameter. On the *conTest* semantic diversity dataset, Confidence NLI Diversity achieves the highest correlation of all metrics (0.62) and approaches human performance. Baseline NLI Diversity performs comparably to the top-performing automatic metric from Tevet et al. (2021), at 0.59 correlation. Although Neutral NLI Diversity does relatively poorly on *conTest* (0.24), it is the highest-performing NLI metric on *decTest* (0.72), suggesting that incorporating neutral predictions may capture lexical instead of semantic diversity.

We perform experimentation using bootstrapping to determine confidence intervals for *conTest* correlations to the diversity parameter. We sample a dataset of 110 elements

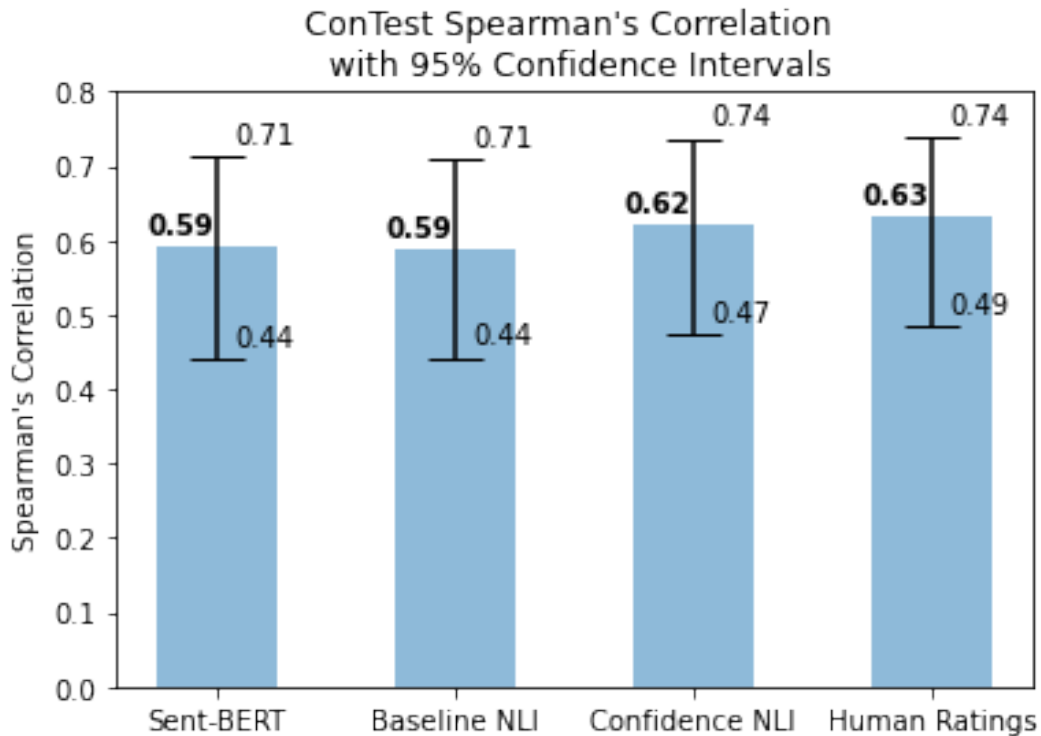


Figure 3.2: Spearman's Correlation with 95% Confidence Intervals.

(50% of the original *conTest* dataset's size) from *conTest* with replacement and compute corresponding Spearman's correlation values using the sampled dataset for Sent-BERT, Baseline NLI Diversity, Confidence NLI Diversity, and human judgments. We repeat this process 1,000 times for stability and calculate 95% Confidence Intervals. The full *conTest* correlation value plotted with these intervals can be seen in Figure 3.2. While the Confidence Interval values overlap between all 4 conditions, the Confidence NLI Diversity distribution closely matches the human distribution.

A histogram of Confidence NLI Diversity values for low and high semantic diversity sets of responses is shown in Figure 3.3. We note the lack of large overlap between the distributions of low and high semantic diversity data. In addition to the correlation results in Sections 3.4 and 3.4, this result indicates the Confidence NLI Diversity metric distinguishes between low and high semantic diversity.

The higher correlation to the diversity parameter leads us to choose NLI Diversity - MNLi instead of Combined for all further experimentation.

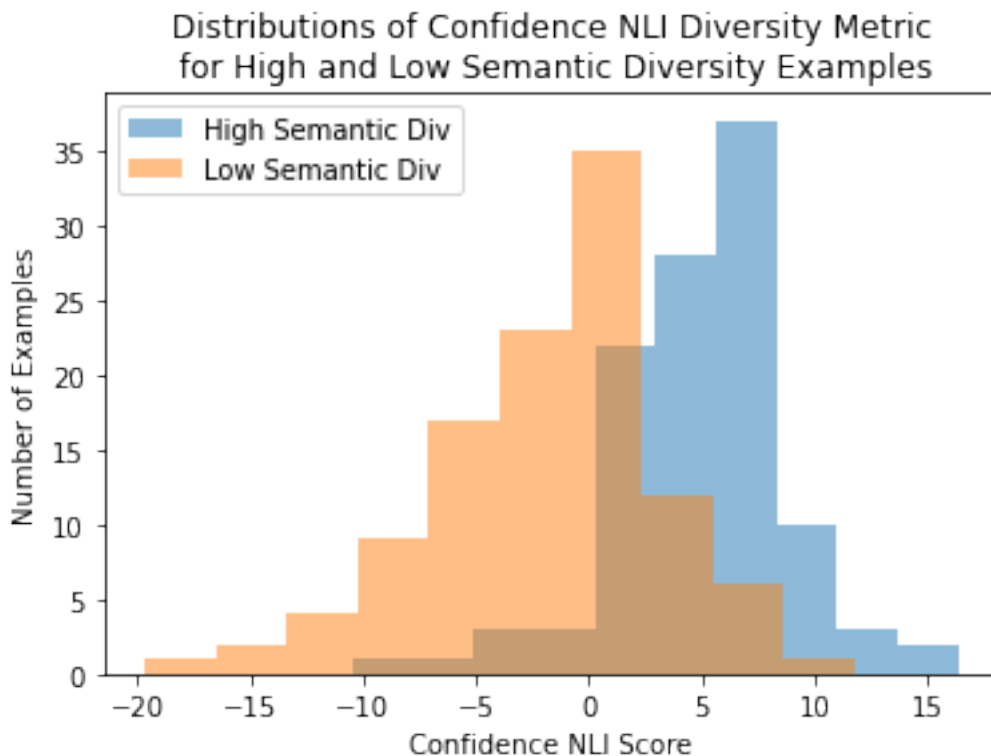


Figure 3.3: Histogram of Confidence NLI Diversity for high and low semantic diversity examples.

Human Correlation

In this subsection, we examine the NLI Diversity metric’s correlation to the human annotations collected by Tevet et al. (2021). Each set of responses in *conTest* and *decTest* is scored by 10 annotators from 1 (not diverse at all) to 5 (very diverse) with half-point increments. We compute correlation with respect to the averaged rating.

In addition to NLI Diversity, we explore the prediction counts for each category. We expect that a higher $\#$ Entailment value will be negatively correlated with diversity because the more pairs of responses that entail each other, the more similar the set of responses is. Similarly, we expect that a higher $\#$ Contradiction value will be positively correlated with diversity. Since the NLI Diversity metric incorporates both $\#$ Entailment and $\#$ Contradiction, we would expect this metric to be highly correlated with human judgments as well.

Spearman’s ρ rank correlation results between our metrics and the human diversity scores are shown in Table 3.3. The highest-performing correlation for lexical diversity is the Neutral NLI Diversity (0.69). The highest-performing semantic diversity correlation is Confidence NLI Diversity (0.64). Additionally, Baseline and Confidence NLI Diversity correlations are stronger when evaluating with the *conTest* dataset than the *decTest* dataset (an increase of

Metric	decTest	conTest
	ρ	ρ
Baseline NLI Diversity	0.48	0.63
Neutral NLI Diversity	0.69	0.40
Confidence NLI Diversity	0.41	0.64
# Contradiction	0.26	0.46
# Neutral	0.05	-0.08
# Entailment	-0.48	-0.65

Table 3.3: Spearman’s ρ correlation between NLI Diversity metrics (MNLI) and human judgments. Negative values indicate higher # Entailment is *negatively* correlated with diversity.

0.48 to 0.63 for Baseline MNLI and 0.41 to 0.64 for Confidence NLI), indicating these metrics are more correlated with human ratings of semantic diversity than lexical diversity.

Across both datasets, # Entailment is negatively correlated with diversity, # Neutral does not have a strong correlation, and # Contradiction is positively correlated, as hypothesized. This supports our motivation to use NLI as a diversity metric.

3.5 Diversity Threshold Generation

We have verified that NLI Diversity is both able to capture semantic diversity and aligns with human judgments. We can additionally use NLI Diversity to define a straightforward desired diversity threshold, div_{thresh} for a set of model-generated responses, u_1, \dots, u_n . For example, we might intend there to be 10 Contradictions within the set. We propose a generation procedure, Diversity Threshold Generation, designed to iteratively increase the diversity of a set of responses for a conversation.

For a conversation, Diversity Threshold Generation begins by sampling n responses. We score the diversity of these responses using a diversity metric, $div_metric(u_1, \dots, u_n)$. If the diversity score falls above div_{thresh} , the process is finished.

If, however, the score falls below div_{thresh} , we identify the model response which contributes *least* to the diversity score by calculating $div_metric(u_1, \dots, u_{n-1})$ for each sub-group of model responses of size $n - 1$. We discard the model response not present in the highest-scoring subgroup and resample a new response. We re-calculate $div_metric(u_1, \dots, u_n)$ and if $div_metric(u_1, \dots, u_n) > div_{thresh}$, the process finishes. We continue resampling until the maximum cutoff of S is reached.

3.6 Evaluation of Diversity Threshold Generation Method

Models and Datasets

We experiment with two neural dialogue models, DialoGPT (700M parameters) (Y. Zhang et al., 2020)⁴ and BlenderBot 1.0 (300M parameters) (Roller et al., 2021)⁵. We use the default Transformers implementation for each model (Wolf et al., 2020) and do not fine-tune them. Runtime was between 3 and 36 hours on one Titan-X GPU.

All experiments involve the dialogue model M generating 5 responses for each conversation. The maximum number of samples, S , is set to 20. All experiments are averaged over 10 trials for stability.

We evaluate each model on the development set of two public English conversational datasets : DailyDialog++ (1,028 conversations) (Sai et al., 2020; Y. Li et al., 2017) and EmpatheticDialogues (2,763 conversations) (Rashkin et al., 2019). DailyDialog++ includes 5 human-written responses per conversation, allowing for multi-reference comparison. We split each EmpatheticDialogues conversation at a random turn (consistent for all experiments) for generation. Since BlenderBot supports up to 128 positional embeddings, we pass in the last 128 tokens of the conversation for this condition.

Metrics

We evaluate three diversity metrics: two semantic diversity metrics, Baseline NLI Diversity (Section 3.3) and Sent-BERT (Reimers et al., 2019; Tevet et al., 2021), and one lexical diversity metric, distinct-n (J. Li et al., 2016a; Tevet et al., 2021). For Sent-BERT, we compute the average negative cosine similarity between BERT sentence embeddings for each pair of responses. Like Tevet et al. (2021), for distinct-n, we compute the average distinct n-grams from $n \in 1, 2, 3, 4, 5$.

Because Baseline NLI Diversity is more human-interpretable than Confidence NLI Diversity, we use this version for experimentation. For all NLI Diversity experiments, div_{thresh} is achieved when # Contradictions is greater than 10 out of a total of 20 pair-wise comparisons. For both Sent-BERT and distinct-n, however, we do not have a human-specifiable threshold. We use empirical thresholds measured from the sets of 5 human responses for each conversation in DailyDialog++. We choose the 90th percentile for div_{thresh} (0.98 and -0.179 for distinct-n and Sent-BERT, respectively).

We decode using nucleus sampling ($p = 0.9$), as it has been shown to increase response diversity (Holtzman et al., 2020). However our method could be applied with other decoding procedures.

⁴https://huggingface.co/transformers/model_doc/dialogpt.html

⁵https://huggingface.co/transformers/model_doc/blenderbot.html

Metric	Model	Dataset	Start- ing Div.	End- ing Div.	Num. Sam- pled
Baseline NLI	DG	Daily	4.11	10.24	6.3
		Emp	3.68	10.11	7.1
	BB	Daily	-5.55	2.51	14.4
		Emp	-8.90	-1.72	16.5
Distinct-n	DG	Daily	0.95	0.98	5.4
		Emp	0.43	0.52	20.0
	BB	Daily	0.61	0.80	20.0
		Emp	0.52	0.71	20.0
Sent- BERT	DG	Daily	-0.26	-0.16	5.2
		Emp	-0.28	-0.16	5.8
	BB	Daily	-0.62	-0.40	19.0
		Emp	-0.71	-0.52	19.7

Table 3.4: Diversity results of using Diversity Threshold Generation (with a div_{thresh} of 10 # Contradictions for NLI, 0.98 for distinct-n, and -0.164 for Sent-BERT). Num. sampled has a maximum value of 20; DG is the DialogGPT model; BB is BlenderBot.

In order to robustly evaluate Diversity Threshold Generation, we measure both (i) whether Diversity Threshold Generation is able to generate more diverse sets of responses than was originally sampled and (ii) whether the increased diversity comes at the expense of decreased *relevancy* of the responses.

Diversity Results

We aim to measure whether the diversity of the 5 responses from M increases using Diversity Threshold Generation, compared to the initial 5 sampled responses. Diversity of the starting and ending sets of utterances is measured by Baseline NLI Diversity, distinct-n, or Sent-BERT. We also report the number of sampled utterances required to reach div_{thresh} .

Results for Diversity Threshold Generation are shown in Table 3.4. For every condition, we see an increase from starting to ending diversity; for NLI Diversity, this results in an average 137% increase. For most conditions, distinct-n requires more samples than Sent-BERT and Baseline NLI Diversity.

We can use the results of Diversity Threshold Generation to probe differences in the models. In our experimental setup, DialoGPT generates more diverse utterances across all

conditions than BlenderBot. The models change by similar proportions from starting to ending diversity using the NLI metric. However, the starting diversity for BlenderBot is far lower than DialoGPT; the negative value for BlenderBot indicates that a large number of entailment predictions were present in the starting response set.

We can also examine differences between the datasets. For instance, we observe lower starting diversities for the Empathetic Dialogues dataset than for DailyDialog++ for both models. Additionally, the number of samples required for EmpatheticDialogues is consistently higher than for DailyDialog++. This is likely because div_{thresh} for both datasets was calculated using human responses from DailyDialog++, since EmpatheticDialogues does not include multiple human responses.

Metric	Model	Dataset	Utterance Overlap
NLI	DG	Daily	2.63
		Emp	2.42
	BB	Daily	1.78
		Emp	1.73
Distinct-n	DG	Daily	2.89
		Emp	0.87
	BB	Daily	1.51
		Emp	1.65
Sent-BERT	DG	Daily	3.11
		Emp	3.0
	BB	Daily	1.56
		Emp	1.64

Table 3.5: Average utterance overlap from starting to ending set of responses using Diversity Threshold Generation on multiple models, datasets, and diversity metrics.

We additionally measure the number of utterances which occur in both the starting and ending sets of responses, called utterance overlap. A high utterance overlap represents a set of responses which did not need to be significantly changed to reach div_{thresh} . For example, an utterance overlap of 4 indicates that only 1 response needed to be resampled (potentially multiple times) from the starting set to reach div_{thresh} . Results are seen in Table 3.5. Keeping in mind that higher Average Overlap indicates less resampling was needed, we note higher overlap for DialoGPT than BlenderBot 1.0 (with the exception of distinct-n and EmpatheticDialogues).

Table 3.6 shows randomly-sampled examples from the DailyDialog++ dataset, created using Diversity Threshold Generation with the DialoGPT model and NLI Diversity as the

intended *div_metric*.

Context	Starting Responses	SD	Ending Responses	ED
<p>Speaker 1: Is this the right counter to check in for this flight?</p> <p>Speaker 2: Yes, it is . Please put your luggage on this scale.</p> <p>Speakaer 1: OK . Is the plane on schedule?</p> <p>Speaker 2: Yes, it is . How many pieces of hand luggage do you have?</p>	<ul style="list-style-type: none"> • I don't understand your question. • I'm worried that airline is cutting a deal, and I'd like to avoid this • I have a box of 100 hand luggage with hand luggage. • I have about 60. • You will be getting a ticket for the trip. 	7	<ul style="list-style-type: none"> • I don't understand your question. • I have a box of 100 hand luggage with hand luggage. • I have about 60. • You will be getting a ticket for the trip. • Just one, but I was planning on taking a plane home. 	11
<p>Speaker 1: Betty, would you please read this letter of application I've just written? I'd like to have your opinion.</p> <p>Speaker 2: I'd be glad to tell you what I think.</p> <p>Speaker 1: Good! I'm interested in your advice.</p>	<ul style="list-style-type: none"> • I was in a discussion on the internet with a bunch of people. • I've sent you a private message, so if you have any questions or feel like you could use the resources, feel free to send them my way. • Are you an American citizen? • If you have any specific questions about it, I'm happy to answer them. • I'll be in touch! 	3	<ul style="list-style-type: none"> • Are you an American citizen? • I'll be in touch! • Would you like to hear something about the business side of things? • I'd like to recommend you to look at the application, there are lots of important things about it.It.does.nothing to prove anything to you. • You are so adorable. I love that. 	10

<p>Speaker 1: Okay, I 'll take it . By the way, do you have delivery service?</p> <p>Speaker 2: You don 't need to carry the furniture home because we have delivery service.</p> <p>Speaker 1: Do you charge for delivery?</p>	<ul style="list-style-type: none"> • What do you have to worry about? We deliver everything. We take a percentage for any delivery charges. • You can have free shipping if you want • I have a 5 dropoff fee that I pay for. • In the beginning • No. We just get it for free.So, no. 	<p>9</p>	<ul style="list-style-type: none"> • What do you have to worry about? We deliver everything. We take a percentage for any delivery charges. • You can have free shipping if you want • I have a 5 dropoff fee that I pay for. • No. We just get it for free.So, no. • No, there is no charge. I have never actually met the person who bought it. 	<p>14</p>
<p>Speaker 1: Am I too abrupt? Did I scare you? Or you don't want ...</p> <p>Speaker 2: No! I have been expecting this moment for a long time . I mean my answer is' yes'.</p> <p>Speaker 1: Oh, I'm so happy . I will try my best to love you and protect you . I promise.</p>	<ul style="list-style-type: none"> • Aww, stop, we both know you're going to try to love me, you know? • Thanks • I'm blushing. I love you. Thank you. • Yayy, you are making me love you. • I love you. It's a good feeling. 	<p>2</p>	<ul style="list-style-type: none"> • Aww, stop, we both know you're going to try to love me, you know? • I love you. Thank you, your a nice man. • Thank you so much. It's okay. I think you are a good woman. You are awesome. • And I will try. And you are my best friend. Thank you, Sarah. • You can have it all 	<p>10</p>

<p>Speaker 1: Hello? This is Mike here . May I have a word with Grace?</p> <p>Speaker 2: She ' s not at her desk at the moment . Is there any message I can give her?</p> <p>Speaker 1: Yes . Please tell her that tickets to Phantom of the Opera will be delivered in two days . My number is 6521684 0, extension 236.</p> <p>Speaker 2: Ok . I ' ll let her know as soon as possible.</p>	<ul style="list-style-type: none"> • I replied to the ticket. Hope it is still available. • Thank you! May I please have the same? • Will you guys be playing the show again for the next three days? • Thank you. She is relieved now. • I am at my desk. 	<p>4</p>	<ul style="list-style-type: none"> • Thank you! May I please have the same? • Will you guys be playing the show again for the next three days? • Thank you. She is relieved now. • I am at my desk. • No, I m not going to be your wife. I am sure her voice is wonderful. 	<p>10</p>
---	---	----------	---	-----------

Table 3.6: Starting and ending responses with corresponding NLI Diversity for DialoGPT, randomly sampled from the DailyDialog++ dataset. SD = Starting NLI Diversity; ED = Ending NLI Diversity.

Additionally, we investigate the stability of Diversity Threshold Generation by measuring the number of samples required before reaching div_{thresh} across multiple runs of the experiment. We present results for NLI Diversity, DailyDialog++, and DialoGPT and observe similar trends across all other conditions.

Figure 3.4 reports the number of resampled utterances required before reaching the intended number of contradictions. Each bar color represents a different run of the experiment. We do not observe a large difference in number of resamples required between runs of the same condition, indicating that the method is stable. The last bucket contains sets of responses which reached the maximum number of samples, $S = 20$, indicating div_{thresh} could not be reached.

Finally, we evaluate beam search’s ability to generate diverse utterances using Diversity Threshold Generation for DailyDialog++ and NLI Diversity. To compare nucleus sampling to beam search, we generate 25 beams and consider these responses from most to least probable, i.e. if the 5 most likely beams do not satisfy the diversity threshold, we remove the lowest-scoring beam and replace it with the 6th most likely beam. We find the starting NLI Diversity for beam search is -5.05, the ending diversity is 5.35, and an average of 10.97 sampled utterances is required. While the NLI Diversity does improve from the starting to

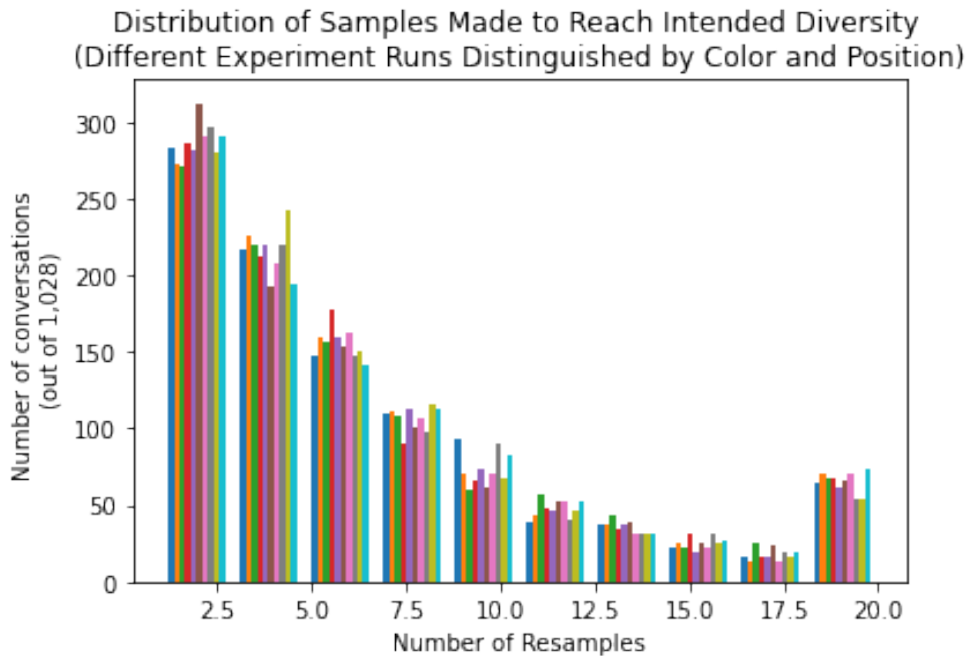


Figure 3.4: Histogram of number of samples required before reaching intended number of contradictions. Each bar color represents a different run of the experiment.

ending set of responses, beam search has a much lower ending diversity than nucleus sampling. While past work has confirmed that nucleus sampling is more *lexically diverse* than beam search using Self-BLEU (Holtzman et al., 2020), our results confirm that nucleus sampling is also able to generate more *semantically diverse* utterances.

Relevance Results

Since past work has documented a tradeoff between diversity and relevancy (Y. Zhang et al., 2018), we also report results for the *relevancy* of the starting and ending sets of responses for Diversity Threshold Generation. We use two established relevancy metrics: BLEU Score (Papineni et al., 2002)⁶ and BERTScore (T. Zhang et al., 2020)⁷. We show results on DailyDialog++, which has multiple human-generated responses for comparison, which is more correlated to human judgments than single-reference evaluation (Gupta et al., 2019).

Results are shown in Table 3.7. The key takeaway is that the relevancy values remain virtually unchanged when using the Diversity Threshold Generation procedure, according to both BLEU score and BERTScore. The average percent difference is 0.08% for BertScore and 1.1% for BLEU.

⁶https://www.nltk.org/_modules/nltk/translate/bleu_score.html

⁷https://github.com/Tiiiger/bert_score

Metric	Model	Starting BERT Score	Ending BERT Score	Start-ing BLEU	End-ing BLEU
NLI	DG	0.862	0.862	0.317	0.318
	BB	0.868	0.867	0.367	0.368
Distinct-n	DG	0.862	0.861	0.319	0.306
	BB	0.867	0.867	0.366	0.367
Sent-BERT	DG	0.863	0.862	0.318	0.313
	BB	0.868	0.867	0.366	0.366

Table 3.7: Results comparing starting and ending sets of responses from Diversity Threshold Generation to sets of human responses using two relevancy metrics, BERTScore and BLEU score.

3.7 Discussion

Limitations. While NLI Diversity is highly-correlated with human judgments of diversity, it is limited by the NLI model chosen. Compared to Sent-BERT, the dataset used to train the NLI model is limited in scope. While our experiments showed that an NLI model trained on more datasets (Combined) did not perform better than MNLI, future work can more explicitly explore the effect of more generalized data on NLI Diversity.

This work is limited by automatic evaluation metrics for diversity and relevance. Future work should conduct additional human validation of model responses. More work could also be done to examine cases where the model was not able to generate diverse set, such as when humans also find creating a diverse set of responses difficult.

Future Work. Our results showed Confidence NLI Diversity was highly correlated with both human judgments and the diversity parameter, achieving state-of-the-art performance on a semantic diversity dataset. The ablation study deepened this finding, showing that NLI contradiction predictions are especially correlated with diversity. Future work can leverage this finding, e.g., by wording crowdworker instructions to ask for generation *contradictory*, rather than *diverse*, responses.

Our results also show that dialogue generation models are able to improve the diversity of a sampled sets of responses using Diversity Threshold Generation. Diversity Threshold Generation can be used to evaluate future models’ capacity to generate multiple diverse responses.

Future work should compare the resulting diverse responses in a conversational context. Studies could be conducted where chatbot users or dialogue writers can choose the way they want the model to respond, similar to Clark et al. (2021).

3.8 Conclusion

We propose a novel semantic diversity metric, NLI Diversity, which is highly correlated to human judgments. Confidence NLI Diversity achieves state-of-the-art results on measuring semantic diversity. We propose Diversity Threshold Generation to incentivize production of diverse sets of responses for a conversation. This results in more diverse sets of responses than originally sampled for multiple models, datasets, and metrics while maintaining relevancy, and can also be used to investigate a model’s ability to produce diverse responses.

Chapter 4

Pragmatically Appropriate Diversity for Dialogue Evaluation

4.1 Introduction

While many different utterances can continue a conversation, some types of responses are more appropriate than others. In contrast to Chapters 2 and 3, which aim to improve the diversity of a model’s responses, this chapter explores the assumption, based on the linguistic pragmatics literature, that the speech acts associated with a conversation can constrain the responses which are appropriate. In particular, some speech acts are part of *adjacency pairs*, where a speaker produces a speech act utterance which is expected to be immediately followed in the conversation by the other speaker’s paired speech act utterance (Levinson, 1983). For example, a Closing utterance by one speaker (such as “Goodbye”) is typically followed by another Closing utterance from the other speaker (Schegloff et al., 1973).

Speech acts can constrain a conversation even if not part of a strict adjacency pair. In relation to how speech acts influence conversational structure, Searle et al. (1985) state:

The key to understanding the structure of conversations is to see that each illocutionary act creates the possibility of a finite and usually quite limited set of appropriate illocutionary acts as replies. Sometimes the appropriate illocutionary act reply is very tightly constrained by the act that precedes it, as in question and answer sequences; and sometimes it is more open, as in causal conversations that move from one topic to another ... each illocutionary act in a conversation creates and constrains the range of appropriate illocutionary responses.

Our work explores whether this limited set of reply actions affects the linguistic diversity of a set of responses to a conversation.

Consider the two conversations presented in Figure 4.1 from DailyDialog++, a conversational dataset with multiple human-provided responses (Sai et al., 2020; Y. Li et al., 2017). We would expect the linguistic diversity of a set of responses for Conversation A to be

Conversation A

Speaker 1: Are you free tonight?

Writer-Generated Responses

- No, not this evening. Can we try for tomorrow night?
- Let's see, I'm free around 8pm. Will that work?
- Yes! What are we getting into?
- Well that depends on what you have planned.
- Didn't we talk about this already? I have a work event tonight.

Conversation B

Speaker 1: Please come in and sit down. I'm happy to finally meet you.

Speaker 2: Same here, Ms. Drake. I've been looking forward to this.

Speaker 1: I'm sorry I kept you waiting.

Writer-Generated Responses

- no problem at all!
- no worries. how are you?
- I was just reviewing some files so it's not a problem.
- you are just in time!
- Let's get started.

Figure 4.1: Two conversations from DailyDialog++ with sets of responses from writers (Section 4.4). When considering different ways the speaker could continue the conversation, the Pragmatically Appropriate Diversity is higher for Conversation A (Yes-No Question) than for Conversation B (Apology).

higher than Conversation B. Why is this the case? Conversation B contains a constraining utterance as the most recent turn in the conversation: "I'm sorry I kept you waiting." Speech act constraints dictate there are a limited number of ways to respond to this utterance and remain pragmatically appropriate. In fact, the first four writer-generated responses to this conversation include some form of Acceptance or Minimization. On the other hand, Conversation A ends with a question. Although the question constrains the next response to

contain an answer, the content of the answer can be expressed in different ways. This can inspire a speaker to create one of many diverse responses to continue the conversation.

Currently, evaluations of chitchat dialogue models do not incorporate these constraints explicitly. Because dialogue models struggle to produce interesting responses, models are evaluated on their ability to produce diverse responses, which can be measured using a variety of different automatic metrics (J. Li et al., 2016a; Tevet et al., 2021; Reimers et al., 2020; Cao et al., 2017; Zhu et al., 2018; Larson et al., 2019), but these do not vary given the conversation’s properties. For this reason, a model’s output would be evaluated the same way whether in response to Conversation A or Conversation B from Figure 4.1.

To address this deficiency, we introduce the notion of Pragmatically Appropriate Diversity (PA Diversity), defined as the extent to which a conversation creates and constrains the creation of multiple diverse responses. To explore this concept systematically, we conduct two analyses. First, we examine both human-labeled and automatically-labeled speech acts of a human-generated multi-response conversational dataset. We hypothesize that the most recent speech act utterance constrains the diversity of responses. Using automatic diversity metrics, we find that the resulting diversity of the human-generated set of responses is significantly different depending on the most recent speech act utterance.

We next propose a novel human evaluation task: rating the extent to which a conversational prompt inspires the creation of multiple diverse responses. Past work in evaluating creative products finds that experts are reliably able to judge the quality of creative products; however, non-expert participants are not able to do this (Amabile, 1982; Baer et al., 2004). Thus, we choose to use a participant population of creative writers for the task of judging PA Diversity. This study finds significant differences among human PA Diversity ratings for different speech acts. We also find that these differences align with both the measured differences from the multiple human-generated responses and our hypotheses about which speech acts cause more diverse responses. Our findings successfully incorporate pragmatics into diversity-based assessment of neural-based dialogue systems.

4.2 Related Work

We place this chapter in the context of related work in dialogue diversity, pragmatics, and creativity evaluation.

Diversity Metrics for Dialogue

Several automatic metrics have been proposed to measure the diversity of dialogue model responses. One category of metrics assumes the dialogue model generates *one* response for each conversation in the test set (the *Test Set Diversity* setting). A frequently used metric is distinct-n, which measures the number of distinct n-grams across all generated sequences (J. Li et al., 2016a). Other metrics measure the number of distinct responses over the test set

(Cao et al., 2017) or compute BLEU score among the set of model-generated responses (Zhu et al., 2018).

A recently posed alternative measures the diversity of a *set* of model responses (the *Multi-Response Diversity* setting) (X. Zhang et al., 2019; Tevet et al., 2021). This setting measures the model’s ability to generate multiple diverse responses to a single conversation.

In order to measure diversity, we utilize two diversity metrics in the Multi-Response setting. The first metric calculates a diversity score based on a Natural Language Inference (NLI) model’s predictions among pairs of utterances, described in Chapter 3. A contradiction prediction is given weight of 1, neutral is 0, and entailment is -1. The set of scores for all pairwise combinations are summed to produce the final NLI Diversity score.

The second metric uses Sent-BERT (Tevet et al., 2021), which depends on the similarity of sentence-level BERT embeddings (Reimers et al., 2019). We calculate Sent-BERT by computing the average pairwise cosine similarity between representations of dialogue responses, subtracted from 1 to convert the metric from a similarity metric to a diversity measure and averaged over all pairwise similarities for a set of responses.

Crowdworker Diversity Evaluation

Past work has used crowdworkers to indirectly measure diversity of model responses, by having workers rate specificity (J. Li et al., 2016a) or interestingness (See et al., 2019) of responses. Tevet et al. (2021) explore whether crowdworkers can directly judge the diversity in a set of model responses, finding that crowdworkers are able to do this reliably via a Likert scale for semantic diversity but not for lexical diversity.

We draw on this work to explore whether crowdworkers are able to determine the PA Diversity of input conversations, rather than the diversity of a set of model responses. In contrast to past work, we rely on an expert crowdworking population of creative writers.

Conversation Analysis and Speech Acts

We draw from the fields of conversational analysis (Schegloff et al., 1973) and pragmatics (Levinson, 1983), with a focus on speech acts. In particular, we consider the notion of *adjacency pairs*: the combination of two sequential speech acts produced by different speakers, in which the first speech act requires the production of the second (Levinson, 1983). For example, question-answer, greeting-greeting, and apology-minimization are adjacency pairs.

In this work, we make use of the DailyDialog dataset, which includes manually-labeled high-level speech acts, consisting of one of Inform, Question, Directive, or Commissive (Y. Li et al., 2017) (see Table 4.1 for examples). Because the DailyDialog speech acts are general, we also make use of the more fine-grained speech act categories from the Switchboard Dialogue Act Corpus (Jurafsky et al., 1997), which are an extension of the DAMSL coding scheme (Core et al., 1997) (see Table 4.1 for examples). We use a classifier to automatically label conversations from the DailyDialog corpus with finer-grained speech acts.

Evaluating Creativity

For our human study, we draw from related work in the measurement of creativity. In particular, we use the Creative Assessment Technique (CAT), which establishes a framework to assess creative output (Amabile, 1982). This process engages experts from creative fields to independently judge the creativity of different output (e.g., stories, artwork). Usually this framework is deployed to evaluate creative output produced for the same prompt. However, Baer et al. (2004) showed that CAT produces consistent judgments for products generated in response to different prompts.

This line of research inspires our current work in two ways. First, we deploy the use of expert creative writers instead of typical crowdworkers because CAT requires the use of experts. Past work found that replacing expert judgments with nonexperts resulted in vastly different creative judgments (Kaufman et al., 2008).

Additionally, we treat each input dialogue as a creative *prompt* and ask writers to rate the extent the prompt inspires the creation of diverse responses. This task is suggested as future work in Baer et al. (2004), but to the best of our knowledge, we are the first to apply it in this manner.

4.3 Speech Act Analysis

In order to explore whether diversity of sets of responses changes based on the input conversation, we choose the DailyDialog++ dataset, which has two important properties: (i) multiple human responses for each conversational prompt, and (ii) high-level speech acts manually assigned to each utterance. In this section, we first analyze this data using diversity metrics to test hypotheses about PA Diversity. Next, we also assess automatically-assigned finer-grained Switchboard speech acts.

Dataset

The original DailyDialog dataset was scraped from online English learner conversations along with manually-assigned speech act labels (Y. Li et al., 2017). DailyDialog++ augments the original DailyDialog conversations with multiple augmented responses from human annotators, allowing us to explore the *Multi-Response Diversity* setting. We work with the training set, containing 9,000 conversations, each with 5 potential responses.

Analysis: Human-Labeled Speech Acts

We hypothesize that the most recent speech act utterance in a conversation will influence the diversity of the set of 5 responses. We begin by exploring the original DailyDialog labeled speech acts, which fall into one of four categories based on the structure proposed in Amanova et al. (2016): Inform, Question, Directive, and Commissive (Y. Li et al., 2017). Table 4.1 contains examples of each speech act with corresponding definitions.

DailyDialog++ Speech Act Utterance Definitions and Examples

Inform	<i>Definition:</i> “the speaker provides the addressee certain information which he believes the addressee not to know or not to be aware of, and which he assumes to be correct” <i>Example:</i> I could not relocate in the next year, but might be open to it in the future.
Question	<i>Definition:</i> “the speaker wants to know something, which he assumes the addressee to know, and puts pressure on the addressee to provide this information” <i>Example:</i> Is there any extra pay for that?
Directive	<i>Definition:</i> “concerned with the speaker’s wish that the addressee performs an action” <i>Example:</i> Oh, thank you . I want one.
Commissive	<i>Definition:</i> ‘capture the speaker’s commitments to perform certain actions’ <i>Example:</i> You are probably right . I’ll go right now and apologize . I try hard not to be late but it is difficult with Beijing traffic.

SWBD Speech Act Utterance Examples (Selected Subset)

Statement	Non	Actually, Costa Rica isn’t in South America . It’s in Central America.
Opinion		
Yes-No Question		Do you mind if I put my jacket there?
Wh Question		What materials do they need in order to apply the passport?
Action Directive		Let me see . \$200 makes RIB 1,616 . Here is the cash and exchange memo . Please check it.
Statement Opinion		Well, a good teacher makes good students.
Conventional Closing	Clos-	Ok . Goodbye.
Open Question		How about your experience?
Offers, Commits	Options,	Well, I’ll call you immediately for instructions on the matter.
Thanking		Thank you for your encouragement.
Apology		Oh, I am so sorry.

Table 4.1: Speech acts from DailyDialog and SWBD along with a sample conversational utterance from DailyDialog++. DailyDialog++ speech act definitions are quoted from Amanova et al. (2016). For brevity, we only include one turn for each speech act instead of the entire conversation.

DailyDialog speech acts are manually assigned to each turn in the conversation; however, we are most interested in the most recent turn. An assumption that we make throughout

Speech Act	NLI Diversity \uparrow	Sent-BERT Diversity \uparrow	Num. Convos.
Inform	4.43	0.72	3875
Question	5.72	0.71	2724
Directive	5.20	0.72	1845
Commissive	4.29	0.71	748

Table 4.2: Diversity for multi-response sets responding to different speech acts along with the number of conversations. For both metrics, higher values indicate higher level of diversity.

this chapter is that there is one speech act for each utterance, although utterances can in fact have multiple associated speech acts (Levinson, 1983).

Hypotheses

We explore whether the most recent speech act utterance affects the output diversity of the 5 human-generated responses in DailyDialog++. We hypothesize that Question and Inform will produce more diverse sets of responses than Directive and Commissive. Directive and Commissive center around an action, which we expect to constrain the set of responses. Although Question is part of an adjacency pair, we hypothesize the content of the question can allow for more diverse responses and elaboration than Directive and Commissive.

Results

To analyze our hypotheses, we examine the diversity distributions for the five multi-response sets for each of the four speech acts. For each response set, we calculate a diversity score using either NLI Diversity or Sent-BERT Diversity. We report the average diversity scores using both metrics for each speech act in Table 4.2.

While the Sent-BERT diversity scores are similar for all categories, we find larger differences in NLI Diversity between the speech act categories. In particular, Question is the most diverse, as expected. Commissive is the least diverse for both metrics, supporting our hypothesis that it would yield less diversity than Question and Inform. Unexpectedly, Directive is more diverse than Inform.

To test statistical significance, we run a Kruskal-Wallis test (Kruskal et al., 1952) with a Dunn posthoc (Dunn, 1964) and a Bonferroni adjustment to account for multiple pairwise comparisons (Haynes, 2013). For NLI Diversity, we note significant ($p < 0.05$) differences among all pairwise categories except for Inform + Commissive. For Sent-BERT Diversity, we note significant ($p < 0.05$) differences only among Inform + Commissive, Question + Commissive, and Directive + Commissive.

Analysis: Model-Labeled Speech Acts

The speech act categories for DailyDialog++ are human-labeled but quite broad. Since we are interested in investigating more fine-grained categories such as different types of questions, we also consider a speech act set containing 42 fine-grained categories: the Switchboard SWBD-DAMSL speech acts (Jurafsky et al., 1997; Core et al., 1997). Although there are domain differences between DailyDialog++ and Switchboard, both deal with social conversations and we believe have similar speech acts represented.

Classification Model

In order to apply the SWBD classes to the DailyDialog++ dataset, we rely on a speech act classification model. We use the baseline RoBERTa-based classification model from He et al. (2021)¹. Our trained model achieves an accuracy of 0.82 when evaluated with gold speech act labels of the SWBD test set. Additionally, a manual examination of 100 classifications for the subset of speech acts we examine yields 85 correct predictions.

For each conversation in DailyDialog++, we pass the most recent conversational turn to this classifier, obtaining a speech act classification of one of the 42 categories in SWBD. Instead of considering all speech act categories, we restrict our analysis to 10 which are most prevalent in DailyDialog++ (predicted at least 100 times by the classifier across the training set). We exclude the “Continued” speech act from consideration, because it is specific to speech acts which are continued from a past turn in SWBD and is not applicable to the DailyDialog++ dataset.

Hypotheses

We hypothesize that more constraining speech acts, such as apology, thanking, and conversational closing, will have lower diversity. Closing-closing, thanking-acceptance, and apology-minimization are all adjacency pairs with constraining responses (Levinson, 1983). Although question-answer is also an adjacency pair, we expect the answer content to contain more diverse content than closing, acceptance, and minimization. Statements of Opinion and Non-Opinion do not generally constrain responses, so we expect diversity of these responses to vary.

Results

In this subsection we present results for the 10 speech acts described above. The NLI Diversity of DailyDialog++ responses for SWBD categories can be seen in Figure 4.2. The highest diversity is seen in response to Yes-No and Open Questions, as expected. Additionally, as expected, the Thanking and Closing categories result in the lowest NLI Diversity. Unexpectedly, Apology has a higher diversity than both Statement of Opinion

¹<https://github.com/zihaohe123/speak-turn-emb-dialog-act-clf>

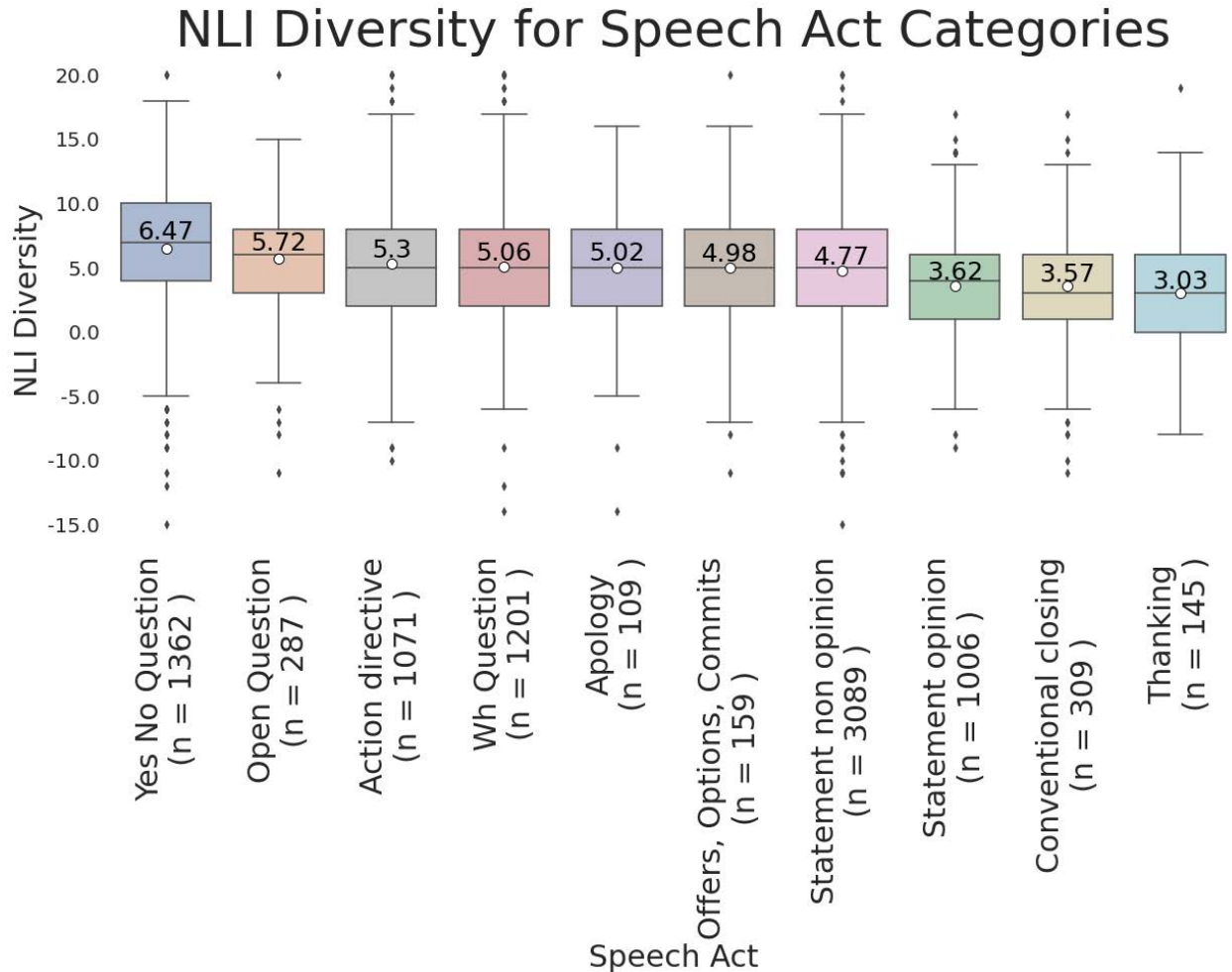


Figure 4.2: NLI Diversity comparison for responses categorized by most-recent speech act utterance (higher values indicate more diverse, ordered by diversity). Mean values are indicated by the white circle and corresponding text label. Box-and-whisker plots show the interquartile ranges.

and Non Opinion; we explore this finding in Section 4.4. It is also surprising that Yes-No Question falls much higher than Wh Question, likely due to the emphasis on contradictions of the NLI Diversity metric.

We report pairwise difference of means with corresponding significance levels for NLI Diversity in Figure 4.3. Looking at a row of this figure, the presence of blue (positive) values indicates the speech act is more diverse; in contrast, the presence of red (negative) values indicates the speech act is less diverse. For example, the “Yes No Question” row has high diversity and the “Thanking” row has low diversity.

Significance for Figure 4.3 is computed with a Kruskal-Wallis test, Dunn posthoc, and

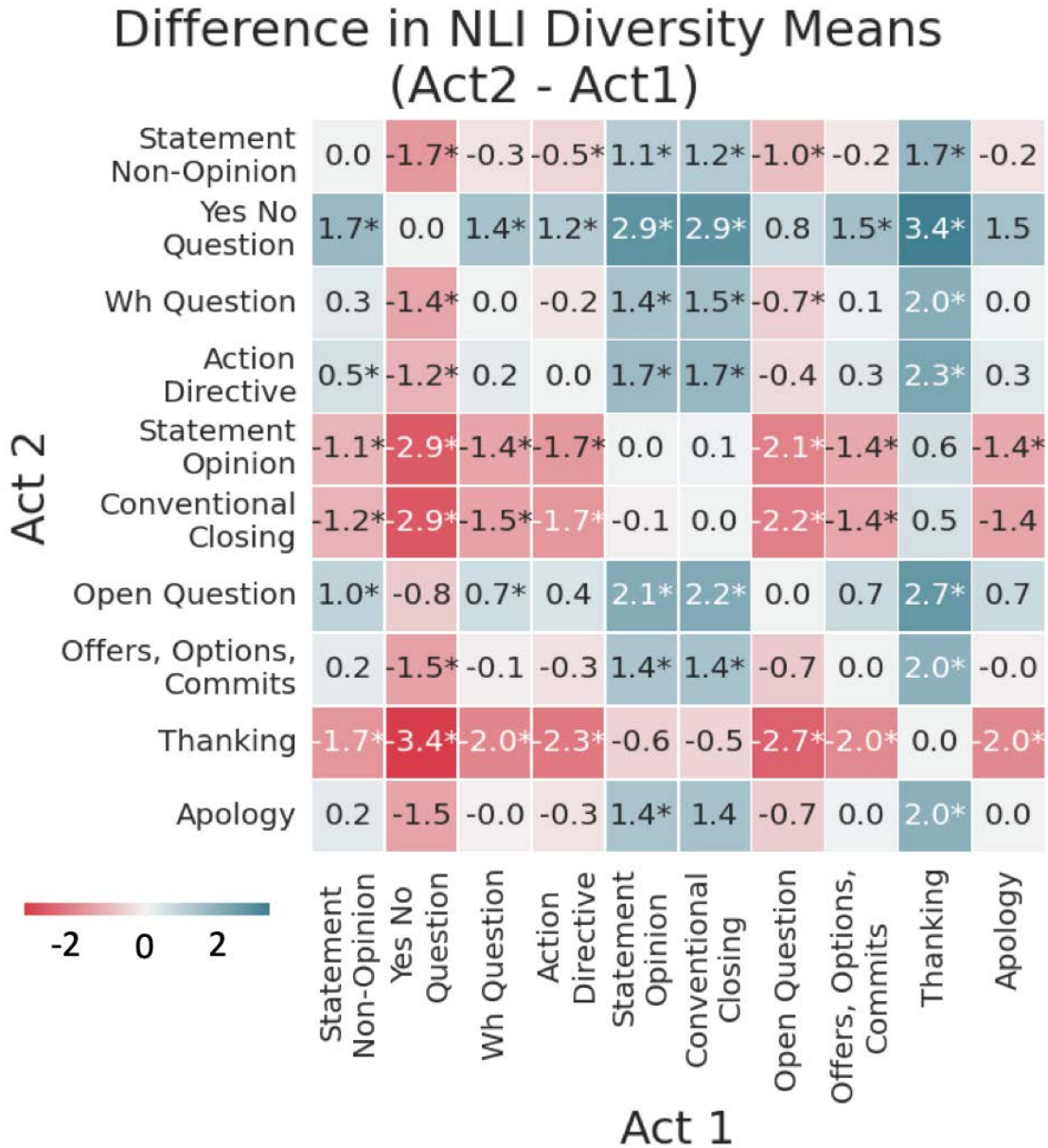


Figure 4.3: Differences in NLI Diversity means between different speech acts. * indicates significance with $p < 0.05$. Darker blue rows indicate the Act2 speech act is more diverse than the Act1 speech act.

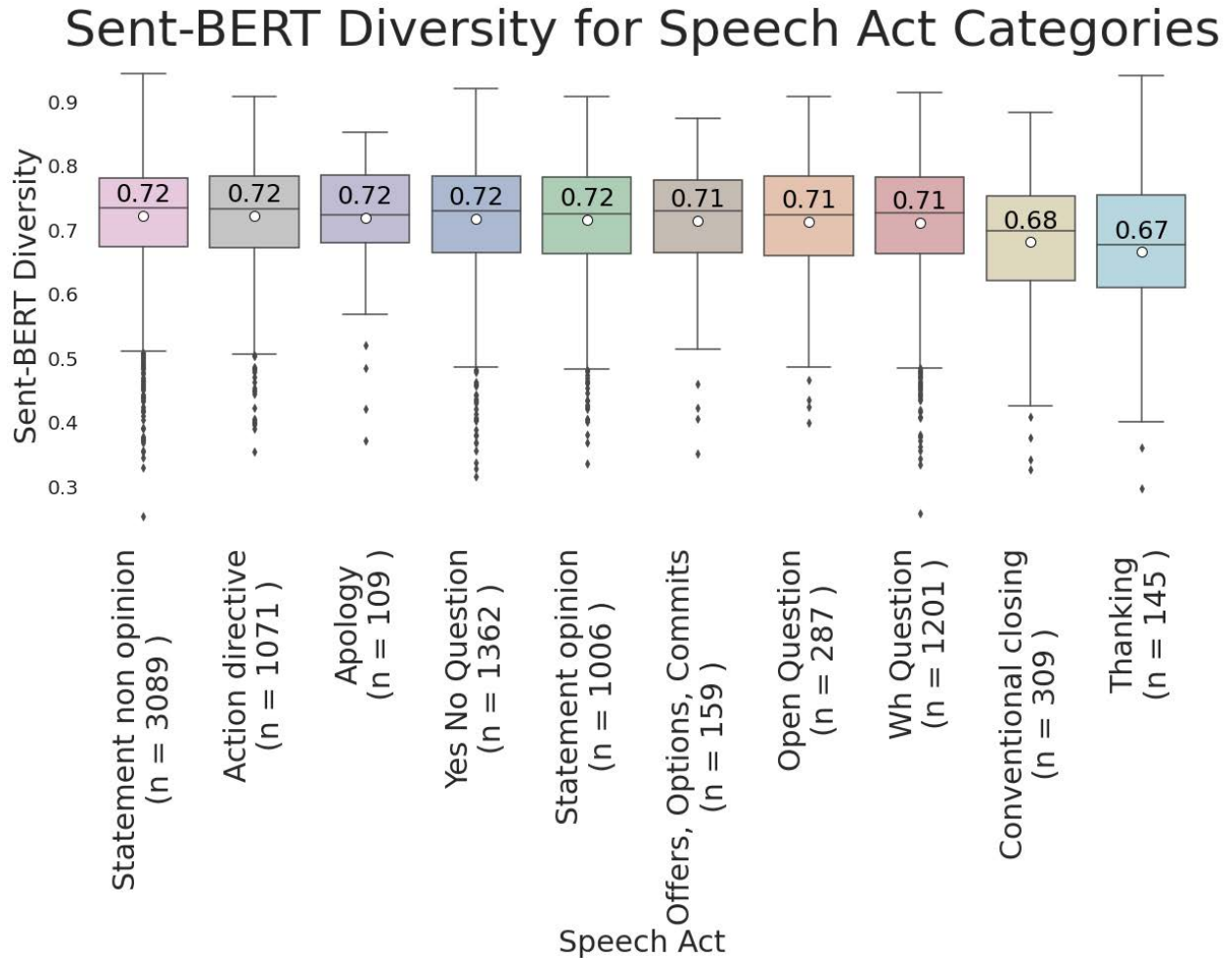


Figure 4.4: Sent-BERT comparison for responses categorized by most-recent speech act utterance (higher values indicate more diverse, ordered by diversity). Mean values are indicated by the white circle and corresponding text label. Box-and-whisker plots show the interquartile ranges.

Bonferroni adjustment, and significant results have $p < 0.05$. We note Yes-No Question, Wh-Question, Action Directive, and Open-Question diversity is significantly higher than most other categories while Thanking and Closing are significantly lower. These significance results support our hypotheses about different diversity scores in response to different speech acts.

We additionally compare Sent-BERT diversity scores to NLI scores to see if the differences in diversity for different speech acts are consistent across multiple diversity metrics. Sent-BERT results for SWBD categories can be seen in Figure 4.4. Sent-BERT scores are more similar across categories, with the exception of Closing and Thanking which are lower.

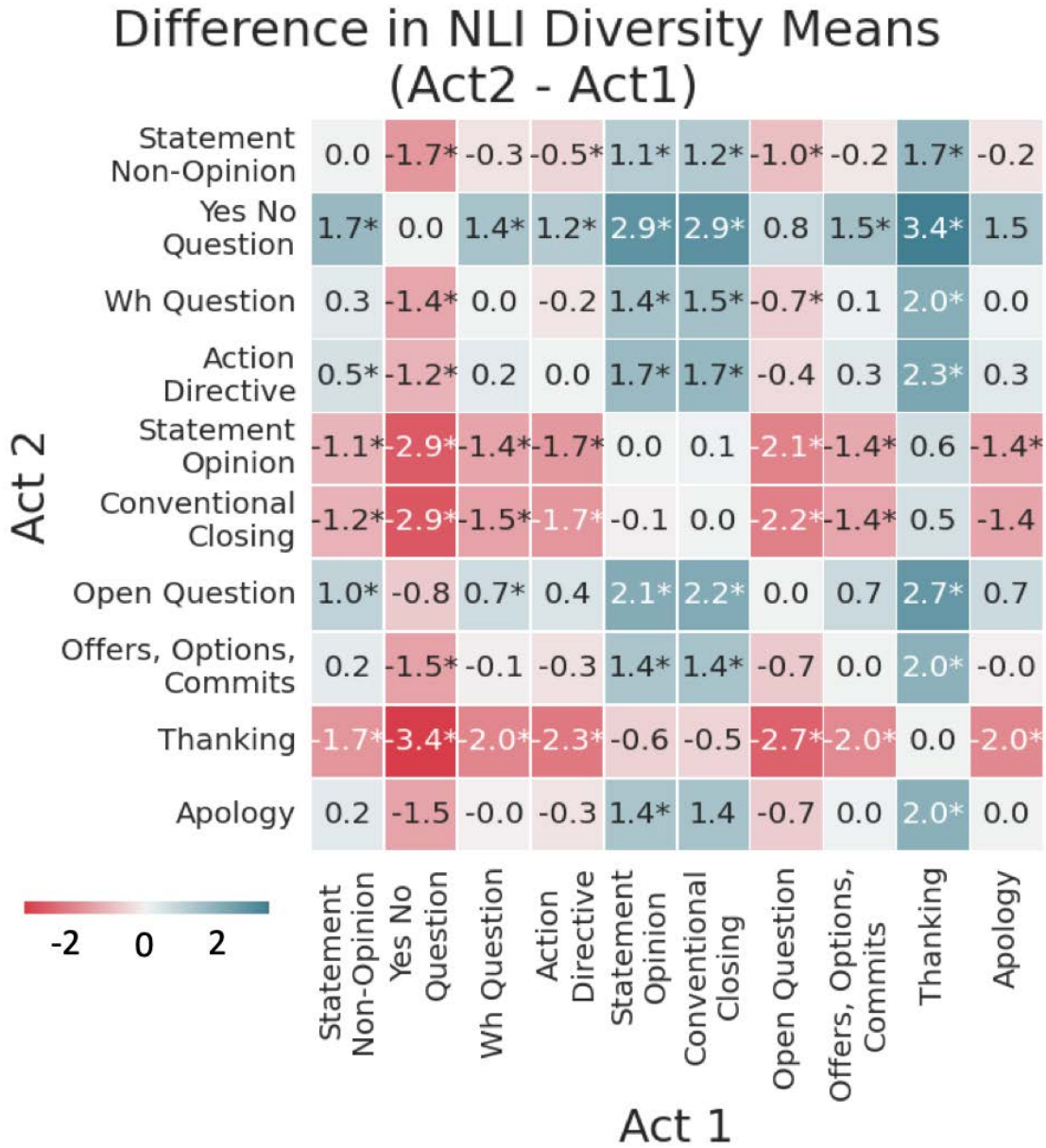


Figure 4.5: Differences in Sent-BERT Diversity means between different speech acts. * indicates significance with $p < 0.05$. Darker blue rows indicate the Act2 speech act is more diverse than the Act1 speech act.

To explore significance among categories for the Sent-BERT comparison, we perform a Kruskal-Wallis test with Dunn posthoc and Bonferroni adjustment. We present mean differences with significance results in Figure 4.5. We note the low (red) rows are Conventional Closing and Thanking, similar to NLI Diversity findings. Overall, we find that Sent-BERT diversity scores give less insight into potential differences between speech act classes than NLI Diversity. Thus, for the next section, we chose to use NLI Diversity to explore the creative writing task.

4.4 Study with Creative Writers

We designed a study to determine whether Pragmatically Appropriate Diversity — the extent to which a conversation constrains the creation of multiple diverse responses — can be determined from the prior conversation using a novel human evaluation task inspired by CAT (Section 4.2).

Study Design

Participants

Because the task involves judging the creativity of input conversational prompts, we followed recommendations of Kaufman et al. (2008) and work with expert writers instead of a typical crowdworking population. To this end, we employed participants who have experience with creative writing, screenwriting, or playwriting from the Upwork platform, prioritizing those who had past Upwork experience or professional writing experience. Throughout all experiments, we worked with a total of 28 Upwork writer participants (two participants repeated one task each on different data). Participants were compensated at a rate of \$20 per survey, estimated to take one hour or less.

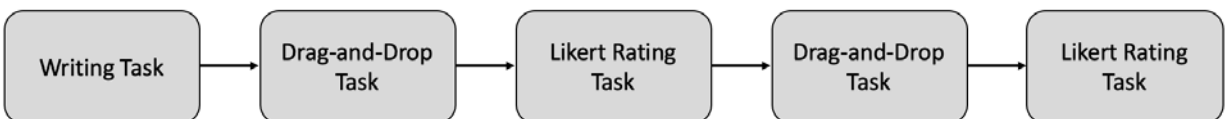


Figure 4.6: Participant progression through sections of the PA Diversity study.

Task Design

A diagram of the section progression can be seen in Figure 4.6. The conditions for each section are:

Writing Task: Participants are instructed to generate “unique, interesting, and appropriate responses to each dialogue conversation.” Participants generate 5 responses each for

Please generate 5 unique, interesting, and appropriate responses to each dialogue conversation.

You should generate these responses as if you are the next speaker responding to the conversation. If the conversation ends with Speaker 1 responding, continue the conversation as if you are Speaker 2. If the conversation ends with Speaker 2 responding, continue the conversation as if you are Speaker 1.

Speaker 1: I think it will take about four hours, including breaks.

Speaker 2: That ' s fine . Are there signposts along the way, in case we get lost?

Speaker 1:

Speaker 1:

Speaker 1:

Speaker 1:

Speaker 1:

Figure 4.7: Writing task from Qualtrics survey.

four conversations (one from each of the four speech acts in the set), presented in a random order. Figure 4.7 shows a screenshot of this section.

Drag-and-Drop Task: This task requires participants to provide a relative ranking for each conversation. Participants are presented with four randomly-ordered conversations (one from each speech act) and asked to drag and drop them such that the top conversation “most inspires the creation of multiple distinct responses” and the bottom conversation “least inspires this.” Figure 4.8 shows a screenshot of this section.

Likert Rating Task: While the drag-and-drop requires participants to rank conversations against one another, we are also interested in their assessments of conversations in isolation. Thus, in this task we ask participants to rate 20 conversations. The rating was on a 5-point Likert scale, where 1 represents “Does not Inspire Creative Responses” and 5 represents “Does Inspire Creative Responses.” We randomize the presentation of the conversations. Figure 4.9 shows a screenshot from this section.

Stimuli Creation

To examine whether the human writings and judgments can uncover differences between different speech acts, we chose two sets of SWBD speech acts with varying levels of NLI Diversity from Section 4.3.

Now, you will rank how well several conversations inspire the creation of several different responses. You do not need to actually generate the responses for these conversations; however, think about the extent to which the conversations inspire you to generate 5 different, creative responses.

Drag and drop each conversation such that 1 is the conversation which most inspires the creation of multiple distinct responses and 4 is the conversation which least inspires this.

Speaker 1: How do you like your name?
Speaker 2: I like it very much . My name has become part of me . I don ' t want to change it.
Speaker 1: What ' s your present (permanent) address?

Speaker 1: I'm sorry but I can't make the final decision myself, and I have to discuss it with other interviewers.
Speaker 2: I know . No matter what the result will be, I have learned a lot from our conversation.
Speaker 1: Your mentality is very good and that's great.
Speaker 2: Thanks for giving me the opportunity of this interview.

Speaker 1: Lucky number . Find its location now.
Speaker 2: We can also consult some from these volunteers.
Speaker 1: Here it is . Well, remember to have our cards on to tell every visitor where we come from.
Speaker 2: Should we put our samples on this table?

Speaker 1: But this place doesn't feel like a forty-ping apartment.
Speaker 2: Well, it's forty pings including the balcony . See . It's huge.
Speaker 1: Oh! The view is great! Mountains, fresh air! Stars?
Speaker 2: This balcony is perfect for barbecuing.
Speaker 1: And we could get a dog!
Speaker 2: I'm sorry . No pets are allowed in this apartment building, except fish.

Figure 4.8: Drag and drop task from Qualtrics survey.

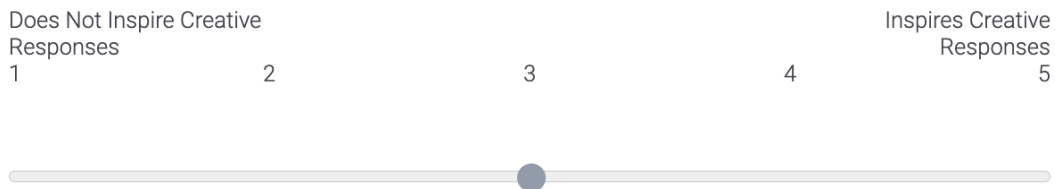
Now, instead of ranking conversations, you will examine several conversations in isolation.

For each conversation, you will assign a score from 1 to 5, based on the extent they inspire creative responses. A score of 1 indicates that a conversation does not inspire the creation of different, creative responses and a score of 5 indicates that a conversation does inspire the creation of different, creative responses.

Speaker 1: What number?

Speaker 2: He's number four.

Speaker 1: Thank you very much for your cooperation.



Speaker 1: Could you help me to register for a class?

Speaker 2: Do you know what class you want to take?

Speaker 1: Are there any Psychology classes available?

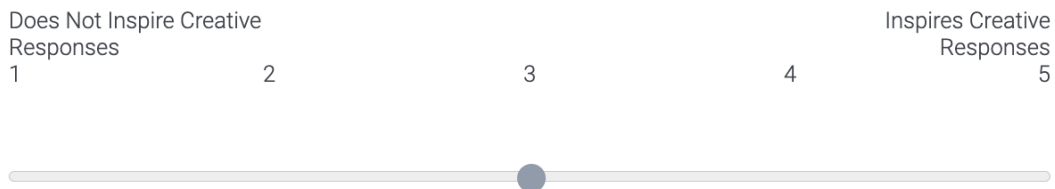


Figure 4.9: Likert rating task from Qualtrics survey.

Set 1 consists of: Yes-No Question, Wh Question, Thanking, and Apology. We hypothesize that the Question acts will result in higher PA Diversity ratings while Thanking and Apology will result in lower ratings, based on trends observed in Section 4.3. This set is chosen to display the largest predicted differences between speech acts.

Set 2 consists of: Open Question, Opinion Statement, Non-Opinion Statement, and Closing. We hypothesize that Open Question and Opinion will have higher diversity ratings and Non-Opinion and Closing will have lower ratings. However, these speech acts were not the most or least diverse. Therefore, we expect these rating differences to be less pronounced than *Set 1*'s.

We intentionally selected conversations whose NLI Diversity scores fell into the median range for the speech act, to ensure the data collected was from prototypical speech act conversations, not outliers. In cases for which sampling conversations from only the median would not result in enough conversations for all surveys, we increased this window to include ± 3 around the median NLI Diversity value. We manually verified the most-recent turn was classified in the correct speech act category, removing misclassifications.

Study Conditions

Each participant completed 52 activities on different conversations, which were evenly distributed among 4 speech acts (see Figure 4.6). We constructed 6 Qualtrics surveys which collectively covered 156 conversations, or 39 conversations per speech act. We recruited five participants for each survey, resulting in 30 completed surveys.

Study Results

This section presents results for the responses by creative writer participants. For the writing task, we find differences that align with our hypotheses; however, significance results are limited. For both the Drag-and-Drop comparison tasks and the Likert ratings, we find significant differences for *Set 1*, which align with our hypotheses. Results for *Set 2* generally support the hypotheses, with some caveats.

Writing Task

We begin by measuring the diversity of the sets of 5 responses produced in the Writing component of the survey. Table 4.3 reports the average NLI and Sent-BERT diversity of these produced responses. Our analysis with responses produced by creative writers replicates the results we saw with the DailyDialog++ responses: For *Set 1*, Yes-No and Wh are more diverse than Apology and Thanking. For *Set 2*, Open Question is more diverse than Closing, also as expected. Unexpectedly, Non-Opinion and Closing are more diverse than Opinion using NLI Diversity, but the difference is relatively small.

For the NLI Diversity of the *Set 1* responses, we find significant results ($p < 0.05$) using Kruskal-Wallis with a Dunn posthoc with Bonferroni adjustment between Yes-No Question

Speech Act		NLI Diversity ↑	Sent-BERT Diversity ↑
Set 1	Yes-No Question	7.3	0.74
	Wh Question	4.8	0.73
	Apology	2.3	0.65
	Thanking	2.6	0.66
Set 2	Open Question	6.2	0.77
	Opinion	3.9	0.74
	Non-Opinion	6.0	0.74
	Closing	4.7	0.74

Table 4.3: Average NLI and Sent-BERT Diversity among sets of participant-produced responses for speech act categories. Speech acts are ordered within each set based on our hypothesized diversity level, from most to least diverse.

+ Apology. Although a Kruskal-Wallis test yields $p < 0.05$ for Set 1 + Sent-BERT, a Dunn posthoc with Bonferroni adjustment does not yield significant pairwise results. For Set 2, for both metrics, the initial Kruskal-Wallis test did not show significance so we did not compute pairwise significance results.

Table 4.4 includes selected writer responses for each speech act. We selected one writer’s responses to show for each set of speech acts, to directly compare how the diversity of one writer’s responses might change based on the most recent speech act utterance.

Speech Act	Prior Conversation	Writer Responses
Set 1		
Yes-No Question	<p>Speaker 1: Honey, where did you buy this skirt?</p> <p>Speaker 2: In the market . Any problem?</p>	<ul style="list-style-type: none"> • No, it’s just very unique and I had to say something about it! • Do you think there’s a problem with it? • Oh, that makes sense. It has some loose threads. I was wondering if it was a thrift item. • Nope, I was just making small talk. I’m a bit frazzled today and I can’t think of much to say. • Yes, it doesn’t go well with your complexion. Perhaps you should find something in a different color.

<p>Wh-Question</p>	<p>Speaker 1: Do you want a taxi? Speaker 2: Yes. Speaker 1: Where to?</p>	<ul style="list-style-type: none"> • I'm visiting the city with a friend but their flight doesn't arrive until tomorrow so I'm sightseeing solo today. What's your favorite area of the city? • I'm headed to work but first I need to stop at the nearest coffee shop. • The nearest nature park, please. I'm in desperate need of fresh air! • I'm not sure. I want to but a few books today but I'm new to the area and wanted to ask a local's opinion on the best used bookstore. I don't want to fund any large corporations. Do you know of a place? • The courthouse. I have a fine to pay. Apparently, it's illegal to leave a couch out on the curb for collection with the garbage. Who knew?!
<p>Apology</p>	<p>Speaker 1: Please come in and sit down . I ' m happy to finally meet you. Speaker 2: Same here, Ms . Drake . I've been looking forward to this. Speaker 1: I ' m sorry I kept you waiting.</p>	<ul style="list-style-type: none"> • No worries! It's hectic times these days! • I understand completely. Please, don't feel guilty about rescheduling. • It's no bother. I'm busy too but I always have time for you! • Not at all. Now, what would you like to discuss? • Admittedly, it wasn't a great first impression, but you can make up for that now. What about ordering some lunch for the two of us?

Thanking	<p>Speaker 1: Would you care to make another appointment?</p> <p>Speaker 2: Unfortunately, I'm leaving town rather unexpectedly, and I may be gone for several days.</p> <p>Speaker 1: I see . I can tell Mr.Thomas that you phoned.</p> <p>Speaker 2: Thank you.</p>	<ul style="list-style-type: none"> • Is there any other message that you would like me to relay to Mr. Thomas? • You are welcome. Will that be all? • Have a nice day, goodbye. • Of course, have a safe trip! • Would you like to get a call back from Mr. Thomas?
Set 2		
Open Question	<p>Speaker 1: How often?</p> <p>Speaker 2: Three times a year at most . How do you feel about this?</p>	<ul style="list-style-type: none"> • That's fine with me. • That sounds a little excessive. How about two? • Only three times a year!?! Sign me up! • I don't think your parents need to come into town that frequently. An annual visit around the holidays should suffice. • It really doesn't matter to me. Whatever you think is best.
Opinion	<p>Speaker 1: I'm looking forward to our son's graduation this weekend.</p> <p>Speaker 2: Yes . So am I . But what will he do after graduation? He really needs to go to college.</p> <p>Speaker 1: Well, dear, we can't force him to go to college . It's up to him.</p>	<ul style="list-style-type: none"> • I understand, but as a parent, I can't help but worry. • True, but I think he'll have better career opportunities if he goes. • What do you mean? I thought we were on the same page about this. • I know. I just want him to do whatever makes him happy. • You're right. Honestly, I'm just ready to have the house back to ourselves.

<p>Non-Opinion</p>	<p>Speaker 1: We scored six goals. Speaker 2: I don't believe it. Speaker 1: And I score three of them. Speaker 2: You are not serious? Speaker 1: We had extra time of course.</p>	<ul style="list-style-type: none"> • And how many points did the other team have? • That's a big improvement! I see all the extra practice is paying off. • How much extra time did you have? • It sounds like it was a great game. I'm sorry I missed it. • Finally! You guys have been on a real losing streak.
<p>Closing</p>	<p>Speaker 1: I just dropped in to say good-bye. Speaker 2: What time are you leaving? Speaker 1: I'm going to try to leave by ten. Speaker 2: Take care and give my best to your parents. Speaker 1: Good-bye . Hope to see you soon again next year.</p>	<ul style="list-style-type: none"> • Ciao! • I hope so too. Feel free to give me a call over the break. • Goodbye! I'll see you in a few months! • Take care of yourself. • I'll hit you up on Instagram!

Table 4.4: Writer responses created in response to each speech act. To directly compare how a writer's responses might change based on speech act, we selected one writer's responses to Set 1 speech acts (Yes-No Question, Wh-Question, Apology, and Thanking) and a separate writer's responses to Set 2 speech acts (Open Question, Opinion, Non-Opinion, and Closing).

A possible explanation for the lack of significance for Sent-BERT is that conversations were selected based on the median NLI Diversity values of response sets in the DailyDialog++ dataset.

Drag-and-Drop Task

Results from the drag-and-drop section can be seen in Figure 4.10. For Set 1, the highest-ranked speech act is Yes-No Question, followed by Wh-Question, Apology, and Thanking, which confirms our observed NLI Diversity for each speech act from Section 4.3. A Friedman's significance test (Friedman, 1940) with Nemenyi posthoc (Nemenyi, 1963) finds that Yes-No Question, Wh-Question, and Apology are rated significantly higher than Thanking, confirming our hypothesis.

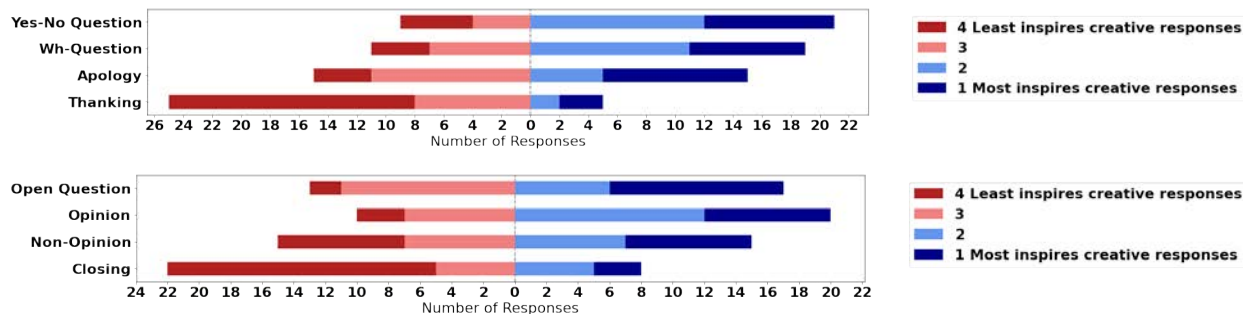


Figure 4.10: Drag-and-drop results for Set 1 (above) and Set 2 (below).

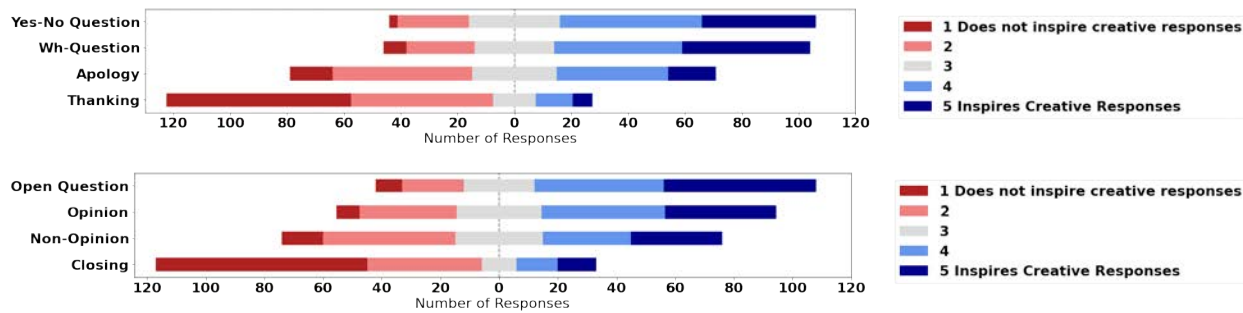


Figure 4.11: Likert results for Set 1 (above) and Set 2 (below).

For Set 2, we observe the Closing speech act results in the lowest ratings. Surprisingly, the Opinion conversations were rated more diverse than the Open Question conversations. Using a Friedman’s significance test with Nemenyi posthoc, however, we only find significant differences between Open Question + Closing and Opinion + Closing. This aligns with our hypothesis that Closing conversations constrain the diversity of responses.

Likert Rating Task

We next examine Likert survey results, first presenting the ratings and then examining whether the ratings correlate to NLI and Sent-BERT diversity scores from DailyDialog++.

Likert Results: Likert ratings for each set of speech acts can be seen in Figure 4.11. For Set 1, we note highest ratings among Yes-No and Wh-Questions, lowest ratings with Thanking, and middle ratings with Apology. This is similar to the drag-and-drop results found in Section 4.4. A Friedman’s significance test with Nemenyi posthoc yields significance ($p < 0.05$) among all pairwise speech acts except for Yes-No + Wh Question.

For Set 2, ratings also aligned with our hypotheses, with Open Question, Opinion, Non-Opinion, and Closing ordered from most- to least-diverse. Contrary to the drag-and-drop results in Section 4.4, we find Open Questions are rated higher than Opinion conversations

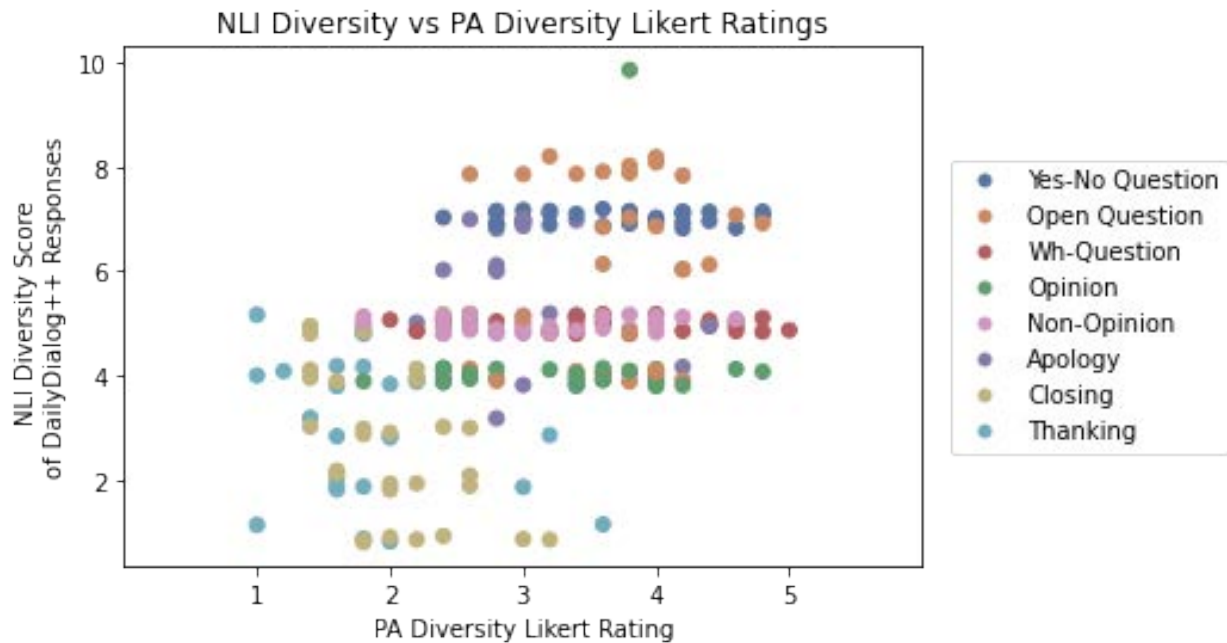


Figure 4.12: Averaged Likert ratings compared to NLI Diversity scores of DailyDialog++ responses. Because NLI Diversity is discrete, we added a random jitter for all points between -0.2 and 0.2.

in the Likert section. We find significant differences ($p < 0.05$) between all pairwise speech act combinations except for Open Question + Opinion and Opinion + Non-Opinion.

Likert Results Compared to DailyDialog++: We next turn to the question of whether participant PA Diversity Likert ratings are correlated to the NLI and Sent-BERT Diversity scores from DailyDialog++ responses. We represent a conversation’s Likert rating by averaging the ratings of the 5 participants who encountered the conversation.

Figure 4.12 plots averaged participant Likert ratings against NLI Diversity scores for corresponding response sets in DailyDialog++. Note that conversations were selected using median values from each speech act category; therefore, this is not a uniform sample of conversations from DailyDialog++ speech acts. The Spearman’s correlation is 0.43 ($p < 0.01$), indicating participant ratings are moderately positively correlated with NLI Diversity scores on the DailyDialog++ data.

A corresponding plot for Sent-BERT diversity can be seen in Figure 4.13. The Spearman’s correlation is 0.19 (very weak, $p < 0.01$), indicating a very weak positive correlation between average Likert ratings and Sent-BERT diversity scores. Future work could investigate if a different conversation sampling strategy would yield a stronger correlation with Sent-BERT diversity scores.

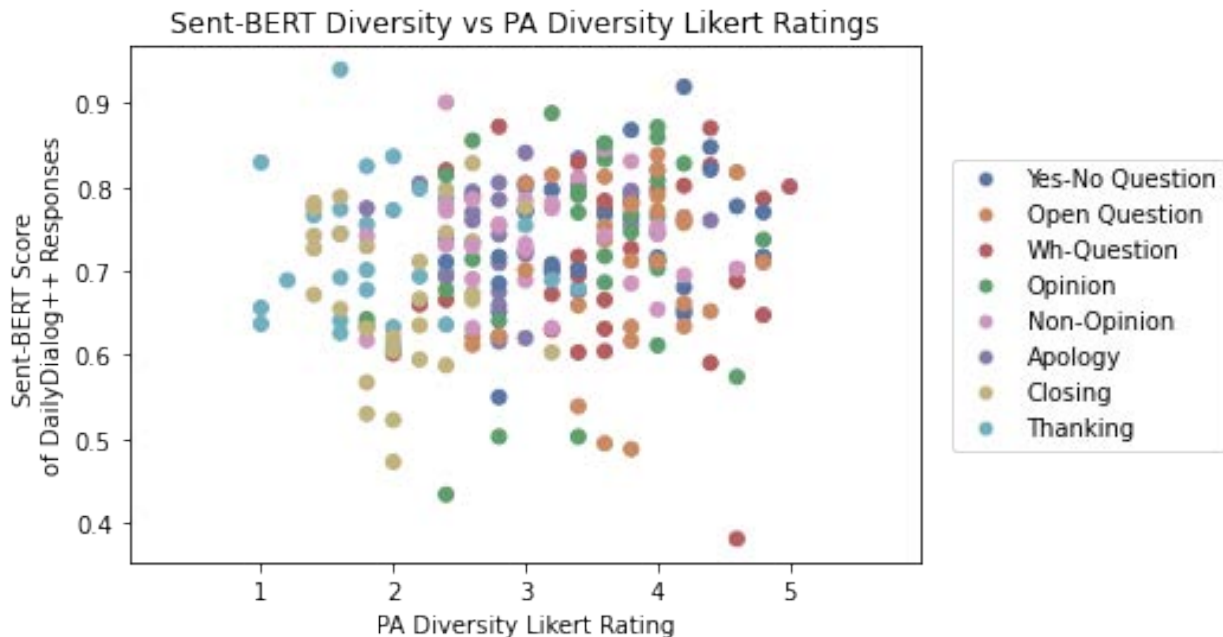


Figure 4.13: Averaged Likert ratings compared to Sent-BERT Diversity scores.

4.5 Discussion and Future Work

We found that controlling for speech act type resulted in significant differences in PA Diversity for sets of responses in DailyDialog++ (Section 4.3) as well as creative writing participant judgments (Section 4.4). In particular, certain types of speech act utterances (Thanking, Apology, and Closing) were consistently judged less likely to produce diverse responses than other speech act utterances.

Implications. These results are significant in their successful incorporation of pragmatics into diversity-based assessment of neural-based dialogue systems. One important implication is that diversity evaluation for dialogue models should be sensitive to the pragmatics of the utterances. This could be done several ways, including:

- Define the PA Diversity for conversations in the evaluation dataset and then assess whether a model’s diversity correlates with PA Diversity scores.
- Report diversity scores separately for high and low PA Diversity conversational turns.

These results have implications for dialogue model generation in addition to evaluation. Perhaps a simpler, rule-based system should be used to generate responses to pragmatically-constraining conversations, relying on a neural network generation system only for conversations with high PA Diversity.

Another important finding of this work is that overall, the NLI Diversity metric, which explicitly incorporates contradiction and entailment predictions, was more distinguishing than the Sent-BERT measure, which incorporates similarity between sentence representations.

This supports results found in Chapter 3.

It is important to note that conversational creativity varies across people; one participant in the study rated every conversation with a score of 5, and wrote: “I tend to think creative options are without limit, especially with limited context.” Thus, when creating and assessing dialogue systems, we should consciously choose if we are creating a witty Oscar Wilde or a conversational partner who follows pragmatic norms.

Future Work. Our findings examined eight different speech acts in depth, but additional common speech acts can be explored, such as Action Directives. While we found significant differences among diversity relating to the most recent speech act utterance, it is possible that the other speech acts in the conversation influence diversity. Future work should also examine the effects of other variables, such as length or the topic of conversation.

To circumvent the need for expert labeling of speech acts, future work should investigate automatically predicting the 5-point PA Diversity rating from expert data. A baseline approach to this task could use the median for the predicted speech act utterance as a proxy.

4.6 Conclusion

We introduce Pragmatically Appropriate Diversity, the hypothesis that dialogue diversity will vary based on the input conversation. An analysis finds significant differences among both human-labeled speech acts from DailyDialog++ and fine-grained speech acts automatically assigned by a SWBD model. We use these findings to create a new human evaluation task, to explore whether the assessments by creative writers align with a conversation’s PA Diversity. We find that writer rankings correspond with our hypotheses, paving the way for PA Diversity to be incorporated into future dialogue evaluation and generation.

Chapter 5

Conclusion

This chapter reflects on the contributions in this dissertation and proposes future work to further the ideas presented.

5.1 Contributions

In the Test Set diversity setting, Chapter 2 proposed a new way of approaching the dialogue diversity problem by collecting a more diverse corpus via DIDC. I found that improving the diversity of the dataset also resulted in increased dialogue model diversity. This finding can inspire work in other areas of the dialogue generation pipeline and can be used to collect corpora with other desired properties.

In the Multi-Response diversity setting, Chapter 3 proposed a new way of measuring diversity using the output of a Natural Language Inference model. This NLI Diversity metric was highly correlated to human judgments, achieved high performance using a diversity evaluation framework, and measured semantic, as opposed to lexical, diversity. This chapter found that *contradictions* were particularly indicative of diversity. Future exploration can further interrogate whether contradictions should be integrated into the instructions for corpus collection. For example, crowdworkers who are identified as less-diverse by DIDC can be instructed to generate contradictory responses, to potentially improve the diversity of their data.

Building on the NLI Diversity metric, Chapter 4 formulated the idea of Pragmatically Appropriate Diversity, the extent to which conversations inspire creation of different diverse responses. This chapter found that the most recent speech act utterance constrained the Pragmatically Appropriate Diversity of human-generated responses, a result which was confirmed with a novel crowdworking writing task. This finding supports the argument that diversity-focused evaluation of dialogue models should explicitly incorporate information about the conversational context. Evaluation should not penalize models responding to conversations in which diversity is constrained by pragmatic appropriateness.

5.2 Future work

One benefit of Diversity-Informed Data Collection in Chapter 2 is that it can be combined with other diversity-increasing strategies. For example, a model can be trained on data collected using DIDC can also utilize diversity-increasing decoding strategies, such as Diversity Threshold Generation in Chapter 3. Both of these techniques can be combined with other diversity-increasing procedures, such as Nucleus sampling (Holtzman et al., 2020). A large scale ablation of different diversity-increasing techniques could provide insight on the benefits of combining different techniques.

Future work should also explore the impact of different types of diversity for different settings. While this dissertation prioritized semantic diversity in Chapter 3, different types of diversity may be desired depending on the application of the chatbot. For example, in a language learning scenario, lexical diversity might be desired to introduce a user to new vocabulary words. While not discussed in this dissertation, syntactic diversity might also be desired, to introduce interesting structure into a conversation or expose a language learner to different grammatical structures.

Additionally, a study can be conducted to measure the effect of a diverse model compare to a baseline model on user engagement in a controlled manner. This study could be conducted in a language learning classroom, where engagement has been linked with learning effectiveness (Carini et al., 2006). Having participants interact with a chatbot for prolonged interactions can provide insights into the real-world effect of a diverse model and can be used to focus future work on types of diversity which are correlated to user engagement.

While this dissertation explored different ways of measuring diversity, future work should explicitly explore which types of diversity these evaluation metrics are capturing. Dialogue responses can be diverse based on many factors, e.g., because they discuss different topics or because speakers are using a different dialect or register. While the Multi-Response setting, where responses are generated for the same conversation, is able to control for conversational topic, future work in dialogue should explore which components are being measured by diversity metrics. Future work could also explore which type of diversity chatbots are better able to capture as well as explicitly modeling these types of diversity, to explore conditioning a chatbot’s generation on different desired properties.

Finally, this dissertation focused on measuring and improving the diversity of *dialogue models*. However, the methods proposed can be used in other areas of NLP. For example, the DIDC method in Chapter 2 can be used in other applications, such as selecting a diverse subset of a dataset to annotate or identifying diverse examples from a dataset to upweight in the training process. The Diversity Threshold Generation procedure proposed in Chapter 3 can be used to improve a desired property in other text generation applications, such as summarization or question generation. And, finally, the findings from Chapter 4 set the stage for explicitly incorporating pragmatic information into other generation tasks.

Bibliography

- Amabile, Teresa M. (Nov. 1982). “Social psychology of creativity: A consensual assessment technique”. In: *Journal of personality and social psychology* 43.5, pp. 997–1013. URL: <https://www.proquest.com/scholarly-journals/social-psychology-creativity-consensual/docview/614292888/se-2?accountid=14496>.
- Amanova, Dilafruz, Volha Petukhova, and Dietrich Klakow (May 2016). “Creating Annotated Dialogue Resources: Cross-domain Dialogue Act Classification”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 111–117. URL: <https://aclanthology.org/L16-1017>.
- Assadi, Sepehr, Justin Hsu, and Shahin Jabbari (2015). “Online Assignment of Heterogeneous Tasks in Crowdsourcing Markets”. In: *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA*. Ed. by Elizabeth Gerber and Panos Ipeirotis. AAAI Press, pp. 12–21. URL: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11598>.
- Baer, John, James C. Kaufman, and Claudia A. Gentile (2004). “Extension of the Consensual Assessment Technique to Nonparallel Creative Products”. In: *Creativity Research Journal* 16.1, pp. 113–117. DOI: 10.1207/s15326934crj1601_11. eprint: https://doi.org/10.1207/s15326934crj1601_11. URL: https://doi.org/10.1207/s15326934crj1601_11.
- Baeza-Yates, Ricardo A., Berthier Ribeiro-Neto, et al. (1999). “Modern Information Retrieval”. In: USA: ACM press New York. Chap. 3, Modeling.
- Baheti, Ashutosh, Alan Ritter, Jiwei Li, and Bill Dolan (Nov. 2018). “Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3970–3980. DOI: 10.18653/v1/D18-1431. URL: <https://www.aclweb.org/anthology/D18-1431>.
- Basu Roy, Senjuti, Ioanna Lykourantzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das (2015). “Task assignment optimization in knowledge-intensive crowdsourcing”. In: *The VLDB Journal—The International Journal on Very Large Data Bases* 24.4, pp. 467–491.
- Batra, Dhruv, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich (2012). “Diverse M-Best Solutions in Markov Random Fields”. In: *Computer Vision –*

- ECCV 2012*. Ed. by Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–16.
- Bederson, Benjamin B. and Alexander J. Quinn (2011). “Web Workers Unite! Addressing Challenges of Online Laborers”. In: *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '11. Vancouver, BC, Canada: Association for Computing Machinery, pp. 97–106. DOI: 10.1145/1979742.1979606. URL: <https://doi.org/10.1145/1979742.1979606>.
- Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (Sept. 2015). “A large annotated corpus for learning natural language inference”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: <https://aclanthology.org/D15-1075>.
- Brawley, Alice M. and Cynthia L.S. Pury (2016). “Work experiences on MTurk: Job satisfaction, turnover, and information sharing”. In: *Computers in Human Behavior* 54, pp. 531–546. DOI: <https://doi.org/10.1016/j.chb.2015.08.031>. URL: <http://www.sciencedirect.com/science/article/pii/S0747563215301072>.
- Cao, Kris and Stephen Clark (Apr. 2017). “Latent Variable Dialogue Models and their Diversity”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 182–187. URL: <https://www.aclweb.org/anthology/E17-2029>.
- Carbonell, Jaime and Jade Goldstein (1998). “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries”. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: Association for Computing Machinery, pp. 335–336. DOI: 10.1145/290941.291025. URL: <https://doi.org/10.1145/290941.291025>.
- Carini, Robert M., George D. Kuh, and Stephen P. Klein (Feb. 2006). “Student Engagement and Student Learning: Testing the Linkages*”. In: *Research in Higher Education* 47.1, pp. 1–32. DOI: 10.1007/s11162-005-8150-9. URL: <https://doi.org/10.1007/s11162-005-8150-9>.
- Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia (Aug. 2017). “SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1–14. DOI: 10.18653/v1/S17-2001. URL: <https://aclanthology.org/S17-2001>.
- Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil (Nov. 2018). “Universal Sentence Encoder for English”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 169–174. DOI: 10.18653/v1/D18-2029. URL: <https://www.aclweb.org/anthology/D18-2029>.

- Chang, Jonathan P., Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil (2019). *ConvoKit: The Cornell Conversational Analysis Toolkit*.
- Clark, Elizabeth and Noah A. Smith (June 2021). “Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 3566–3575. DOI: 10.18653/v1/2021.naacl-main.279. URL: <https://aclanthology.org/2021.naacl-main.279>.
- Core, Mark G. and James F. Allen (Nov. 1997). “Coding Dialogs with the DAMSL Annotation Scheme”. In: *Working Notes of the AAI Fall Symposium on Communicative Action in Humans and Machines*. <http://www.scientificcommons.org/43031741>, pp. 28–35. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.7024>.
- Dhole, Kaustubh and Christopher D. Manning (July 2020). “Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 752–765. DOI: 10.18653/v1/2020.acl-main.69. URL: <https://www.aclweb.org/anthology/2020.acl-main.69>.
- Dunn, Olive Jean (1964). “Multiple Comparisons Using Rank Sums”. In: *Technometrics* 6.3, pp. 241–252. DOI: 10.1080/00401706.1964.10490181. eprint: <https://www.tandfonline.com/doi/pdf/10.1080/00401706.1964.10490181>. URL: <https://www.tandfonline.com/doi/abs/10.1080/00401706.1964.10490181>.
- Fan, Ju, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng (2015). “ICrowd: An Adaptive Crowdsourcing Framework”. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’15. Melbourne, Victoria, Australia: Association for Computing Machinery, pp. 1015–1030. DOI: 10.1145/2723372.2750550. URL: <https://doi.org/10.1145/2723372.2750550>.
- Friedman, Milton (1940). “A Comparison of Alternative Tests of Significance for the Problem of m Rankings”. In: *The Annals of Mathematical Statistics* 11.1, pp. 86–92. URL: <http://www.jstor.org/stable/2235971> (visited on 04/23/2022).
- Fryer, Luke K., Mary Ainley, Andrew Thompson, Aaron Gibson, and Zelinda Sherlock Dutta (May 2017). “Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners”. In: *Computers in Human Behavior* 75. DOI: 10.1016/j.chb.2017.05.045.
- Fryer, Luke K. and Rollo Carpenter (2006). “Bots as language learning tools”. In: *Language Learning & Technology* 10.3, pp. 8–14.
- Fryer, Luke K., David Coniam, R. Carpenter, and Diana Lăpuşneanu (2020). “Bots for Language Learning Now: Current and Future Directions.” In: *Language Learning & Technology* 24, pp. 8–22.
- Fryer, Luke K., Kaori Nakao, and Andrew Thompson (2019). “Chatbot learning partners: Connecting learning experiences, interest and competence”. In: *Computers in Human Behavior* 93, pp. 279–289. DOI: <https://doi.org/10.1016/j.chb.2018.12.023>. URL: <http://www.sciencedirect.com/science/article/pii/S0747563218306095>.

- Ghandeharioun, Asma, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard (2019). “Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.
- Gimpel, Kevin, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich (Oct. 2013). “A Systematic Exploration of Diversity in Machine Translation”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1100–1111. URL: <https://aclanthology.org/D13-1111>.
- Gupta, Prakhar, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham (Sept. 2019). “Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References”. In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics, pp. 379–391. URL: <https://www.aclweb.org/anthology/W19-5944>.
- Hashimoto, Tatsunori B., Hugh Zhang, and Percy Liang (June 2019). “Unifying Human and Statistical Evaluation for Natural Language Generation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1689–1701. DOI: 10.18653/v1/N19-1169. URL: <https://aclanthology.org/N19-1169>.
- Haynes, Winston (2013). “Bonferroni Correction”. In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota. New York, NY: Springer New York, pp. 154–154. DOI: 10.1007/978-1-4419-9863-7_1213. URL: https://doi.org/10.1007/978-1-4419-9863-7_1213.
- He, Zihao, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani (Nov. 2021). “Speaker Turn Modeling for Dialogue Act Classification”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2150–2157. DOI: 10.18653/v1/2021.findings-emnlp.185. URL: <https://aclanthology.org/2021.findings-emnlp.185>.
- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2020). “The Curious Case of Neural Text Degeneration”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=rygGQyrFvH>.
- Jia, Xin, Wenjie Zhou, Xu Sun, and Yunfang Wu (July 2020). “How to Ask Good Questions? Try to Leverage Paraphrases”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6130–6140. DOI: 10.18653/v1/2020.acl-main.545. URL: <https://www.aclweb.org/anthology/2020.acl-main.545>.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan (June 2016). “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association

- for Computational Linguistics, pp. 110–119. DOI: 10.18653/v1/N16-1014. URL: <https://www.aclweb.org/anthology/N16-1014>.
- Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (Apr. 2017). “Bag of Tricks for Efficient Text Classification”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 427–431. URL: <https://www.aclweb.org/anthology/E17-2068>.
- Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca (1997). *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13*. Tech. rep. 97-02. Boulder, CO: University of Colorado, Boulder Institute of Cognitive Science.
- Kang, Yiping, Yunqi Zhang, Jonathan K. Kummerfeld, Lingjia Tang, and Jason Mars (June 2018). “Data Collection for Dialogue System: A Startup Perspective”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. New Orleans - Louisiana: Association for Computational Linguistics, pp. 33–40. DOI: 10.18653/v1/N18-3005. URL: <https://aclanthology.org/N18-3005>.
- Kaufman, James C., John Baer, Jason C. Cole, and Janel D. Sexton (2008). “A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique”. In: *Creativity Research Journal* 20.2, pp. 171–178. DOI: 10.1080/10400410802059929. eprint: <https://doi.org/10.1080/10400410802059929>. URL: <https://doi.org/10.1080/10400410802059929>.
- Ki, Nayoung, Y. J. Cha, and Hea-Suk Kim (2019). “Future English Learning: Chatbots and Artificial Intelligence”. In: *Multimedia-Assisted Language Learning* 22, pp. 32–53.
- Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. URL: <http://arxiv.org/abs/1412.6980>.
- Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush (2017). “OpenNMT: Open-Source Toolkit for Neural Machine Translation”. In: *Proc. ACL*. DOI: 10.18653/v1/P17-4012. URL: <https://doi.org/10.18653/v1/P17-4012>.
- Kobren, Ari, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich (2015). “Getting More for Less: Optimized Crowdsourcing with Dynamic Tasks and Goals”. In: *Proceedings of the 24th International Conference on World Wide Web. WWW ’15*. Florence, Italy: International World Wide Web Conferences Steering Committee, pp. 592–602. DOI: 10.1145/2736277.2741681. URL: <https://doi.org/10.1145/2736277.2741681>.
- Kruskal, William H. and W. Allen Wallis (1952). “Use of Ranks in One-Criterion Variance Analysis”. In: *Journal of the American Statistical Association* 47.260, pp. 583–621. DOI: 10.1080/01621459.1952.10483441. eprint: <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1952.10483441>. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441>.
- Kumai, Katsumi, Masaki Matsubara, Yuhki Shiraishi, Daisuke Wakatsuki, Jianwei Zhang, Takeaki Shionome, Hiroyuki Kitagawa, and Atsuyuki Morishima (2018). “Skill-and-Stress-

- Aware Assignment of Crowd-Worker Groups to Task Streams”. In: *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*, pp. 88–97.
- Larson, Stefan, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars (June 2019). “Outlier Detection for Improved Data Quality and Diversity in Dialog Systems”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 517–527. DOI: 10.18653/v1/N19-1051. URL: <https://aclanthology.org/N19-1051>.
- Levinson, Stephen C. (1983). *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press. DOI: 10.1017/CB09780511813313.
- Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan (June 2016a). “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 110–119. DOI: 10.18653/v1/N16-1014. URL: <https://aclanthology.org/N16-1014>.
- Li, Jiwei, Will Monroe, and Dan Jurafsky (2016b). “A simple, fast diverse decoding algorithm for neural generation”. In: *arXiv preprint arXiv:1611.08562*.
- Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky (Sept. 2017). “Adversarial Learning for Neural Dialogue Generation”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2157–2169. DOI: 10.18653/v1/D17-1230. URL: <https://www.aclweb.org/anthology/D17-1230>.
- Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu (Nov. 2017). “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 986–995. URL: <https://aclanthology.org/I17-1099>.
- Lin, Christopher H., Mausam, and Daniel S. Weld (2018). “Active Learning with Unbalanced Classes and Example-Generation Queries”. In: *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*. Ed. by Yiling Chen and Gabriella Kazai. AAAI Press, pp. 98–107. URL: <https://aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17927>.
- Liu, Tong, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan (2019). “Learning to Predict Population-Level Label Distributions”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW ’19. San Francisco, USA: ACM, pp. 1111–1120. DOI: 10.1145/3308560.3317082. URL: <http://doi.acm.org/10.1145/3308560.3317082>.

- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692*.
- Nemenyi, Peter Bjorn (1963). “Distribution-free multiple comparisons.” PhD thesis.
- Nie, Yixin, Haonan Chen, and Mohit Bansal (2019). “Combining Fact Extraction and Verification with Neural Semantic Matching Networks”. In: *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Nie, Yixin, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela (July 2020). “Adversarial NLI: A New Benchmark for Natural Language Understanding”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4885–4901. DOI: 10.18653/v1/2020.acl-main.441. URL: <https://aclanthology.org/2020.acl-main.441>.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). “BLEU: A Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL ’02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: <https://doi.org/10.3115/1073083.1073135>.
- Rahman, H., S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das (Nov. 2015). “Task Assignment Optimization in Collaborative Crowdsourcing”. In: *2015 IEEE International Conference on Data Mining*, pp. 949–954. DOI: 10.1109/ICDM.2015.119.
- Rashkin, Hannah, Eric Michael Smith, Margaret Li, and Y-Lan Boureau (July 2019). “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5370–5381. DOI: 10.18653/v1/P19-1534. URL: <https://aclanthology.org/P19-1534>.
- Reimers, Nils and Iryna Gurevych (Nov. 2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: <https://aclanthology.org/D19-1410>.
- (Nov. 2020). “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. URL: <https://aclanthology.org/2020.emnlp-main.365>.
- Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston (Apr. 2021). “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 300–325. DOI: 10.18653/v1/2021.eacl-main.24. URL: <https://aclanthology.org/2021.eacl-main.24>.

- Sai, Ananya B., Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra (2020). “Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining”. In: *Transactions of the Association for Computational Linguistics* 8, pp. 810–827. DOI: 10.1162/tacl_a_00347. URL: <https://aclanthology.org/2020.tacl-1.52>.
- Schegloff, Emanuel A and Harvey Sacks (1973). “Opening up Closings”. In: *Semiotica* 8.4.
- Searle, John R. and Daniel Vanderveken (1985). *Foundations of Illocutionary Logic*. Cambridge University Press Cambridge ; New York ; Melbourne. Chap. 1, p. 11. URL: <http://www.loc.gov/catdir/enhancements/fy0904/84004948-t.html>.
- See, Abigail, Stephen Roller, Douwe Kiela, and Jason Weston (June 2019). “What makes a good conversation? How controllable attributes affect human judgments”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1702–1723. DOI: 10.18653/v1/N19-1170. URL: <https://www.aclweb.org/anthology/N19-1170>.
- Serban, Iulian Vlad, Alessandro Sordani, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio (2017). “A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI’17. San Francisco, California, USA: AAAI Press, pp. 3295–3301.
- Shao, Louis, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil (2017). “Generating Long and Diverse Responses with Neural Conversation Models”. In: *CoRR* abs/1701.03185. arXiv: 1701.03185. URL: <http://arxiv.org/abs/1701.03185>.
- Song, Haoyu, Wei-Nan Zhang, Jingwen Hu, and Ting Liu (2020). “Generating Persona Consistent Dialogues by Exploiting Natural Language Inference”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 8878–8885. URL: <https://aaai.org/ojs/index.php/AAAI/article/view/6417>.
- Stasaski, Katherine, Grace Hui Yang, and Marti A. Hearst (July 2020). “More Diverse Dialogue Datasets via Diversity-Informed Data Collection”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4958–4968. DOI: 10.18653/v1/2020.acl-main.446. URL: <https://www.aclweb.org/anthology/2020.acl-main.446>.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., pp. 3104–3112. URL: <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>.
- Tao, Chongyang, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan (July 2018). “Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism”. In: *Proceedings of the Twenty-Seventh International*

- Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, pp. 4418–4424. DOI: 10.24963/ijcai.2018/614. URL: <https://doi.org/10.24963/ijcai.2018/614>.
- Tevet, Guy and Jonathan Berant (Apr. 2021). “Evaluating the Evaluation of Diversity in Natural Language Generation”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 326–346. DOI: 10.18653/v1/2021.eacl-main.25. URL: <https://aclanthology.org/2021.eacl-main.25>.
- Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal (June 2018). “FEVER: a Large-scale Dataset for Fact Extraction and VERification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: <https://aclanthology.org/N18-1074>.
- Weizenbaum, Joseph (Jan. 1966). “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine”. In: *Commun. ACM* 9.1, pp. 36–45. DOI: 10.1145/365153.365168. URL: <https://doi.org/10.1145/365153.365168>.
- Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. URL: <http://aclweb.org/anthology/N18-1101>.
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (Oct. 2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: <https://aclanthology.org/2020.emnlp-demos.6>.
- Yang, Chun-Ju, Kristen Grauman, and Danna Gurari (2018). “Visual Question Answer Diversity”. In: *Sixth AAAI Conference on Human Computation and Crowdsourcing*.
- Yu, Dian, Michelle Cohn, Yi Mang Yang, Chun Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu (Nov. 2019). “Gunrock: A Social Bot for Complex and Engaging Long Conversations”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, pp. 79–84. DOI: 10.18653/v1/D19-3014. URL: <https://aclanthology.org/D19-3014>.

- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). “BERTScore: Evaluating Text Generation with BERT”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- Zhang, Xinyuan, Yi Yang, Siyang Yuan, Dinghan Shen, and Lawrence Carin (July 2019). “Syntax-Infused Variational Autoencoder for Text Generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2069–2078. DOI: 10.18653/v1/P19-1199. URL: <https://aclanthology.org/P19-1199>.
- Zhang, Yizhe, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujuan Li, Chris Brockett, and Bill Dolan (2018). “Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., pp. 1815–1825.
- Zhang, Yizhe, Siqu Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan (July 2020). “DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 270–278. DOI: 10.18653/v1/2020.acl-demos.30. URL: <https://aclanthology.org/2020.acl-demos.30>.
- Zhao, Tiancheng, Ran Zhao, and Maxine Eskenazi (July 2017). “Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 654–664. DOI: 10.18653/v1/P17-1061. URL: <https://aclanthology.org/P17-1061>.
- Zhou, Sharon, Melissa Valentine, and Michael S. Bernstein (2018). “In Search of the Dream Team: Temporally Constrained Multi-Armed Bandits for Identifying Effective Team Structures”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: ACM, 108:1–108:13. DOI: 10.1145/3173574.3173682. URL: <http://doi.acm.org/10.1145/3173574.3173682>.
- Zhu, Yaoming, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu (2018). “Txygen: A Benchmarking Platform for Text Generation Models”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, pp. 1097–1100. URL: <https://doi.org/10.1145/3209978.3210080>.