Automatic Evaluation of Dialogue-Systems Using Neural-Network Methods

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Rostislav Nedelchev
aus
Varna, Bulgarien

Bonn, 2023

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Jens Lehmann
2. Gutachter:	Prof. Dr. Christian Bauckhage
Tag der Promotion:	19.04.2023
Erscheinungsjahr:	2023

Abstract

We usually interact with computers by means of specialized tools that are not as common as the language humans use. This has motivated researchers for already several decades to develop algorithms that enable interfacing with computer systems using natural language. This is especially prominent in recent times with the rise of voice assistants like Apple Siri or Amazon Alexa. However, the research and development of such systems is expensive in terms of human labor. The high expenses are especially prominent for the evaluation of such systems, which are very often evaluated by human annotators as a final stage and based on expensive development.

The focus of this thesis is to support the assessment of dialogue systems by creating automatic tools that support humans. Human conversations involve many intricacies that makes it difficult to develop an algorithm which could reliably but also informatively evaluate them. To put the challenge into context, one should consider the Turing test, which is a method of examination in artificial intelligence (AI) for ascertaining whether a computer is proficient of thinking like a human being. One of its key components is the ability to decide whether a conversation is natural. There are various criteria according to which a dialogue is evaluated, and hence, problems that is suffers from. In this work, we aim to detect of these problems.

In order to emulate human-like intelligence, we stand on the shoulders of techniques in Natural Language Processing (NLP), machine and deep learning (ML, DL). Since we have the goal to reduce human effort in the evaluation of dialogues, we focus on methods that can achieve our goal without the need of additionally annotated data:

1. We apply approaches from various problem domains. The thesis makes use of out-of-distribution (OoD), and anomaly detection approaches to treat low quality or problematic dialogue utterances as "unusual."
2. Despite being researched for a few decades, Language Models (LMs) became popular in the research only in the last few years. In our work, we show that they too can be used to evaluate dialogue quality.
3. Natural Language Processing as a field aims to teach various human-like language skills to computers, e.g. abilities like understanding whether two sentences are similar in meaning or whether a piece of text has a positive or negative sentiment. We show that these skills can be used as indirect indicators of conversation quality.
4. In addition, we show that dialogue systems can be evaluated not by means of reference, but "opinion." In other words, instead of asking them to generate a solution for a problem, we show that you can ask them to evaluate a reference solution and based on develop an understanding about the abilities of a dialogue system.

All of the proposed approaches in this thesis do not make use of supervision for dialogue evaluation. They manage to deliver insights using various perspectives that could potentially complement each other in an overall framework for assessing conversation quality.

Acknowledgements

First and foremost, I would like to express my gratitude to Prof. Dr. Jens Lehmann, who supervised me throughout my PhD journey. He offered me a once in a lifetime opportunity that I would never miss on. Through his motivation, persistence and support, I would not have been able to push through this journey.

In addition, I would like to thank to Prof. Dr. Ricardo Usbeck for being a mentor, who never allowed me to settle for little, and helped me achieve the best possible version of myself and the work I do.

Furthermore, I would like to show my gratefulness to Dr. Debanjan Chaudhuri and Dr. Denis Lukovnikov, who were great sparring partners, and a tremendous source of inspiration. They always widened my perspective in ways I could not imagine.

Moreover, I express my sincere thanks to Line Winterhoff - my partner with whom I could share any journey through life. Without her immense moral support and understanding, I would have not "conquered heights."

Finally, I would like to thanks family and friends for the wholehearted encouragement, which helped me to never feel alone in this adventure.

Contents

CHAPTER 1

Introduction

Cogito, ergo sum
(*Latin*; I think, therefore I am)

René Descartes

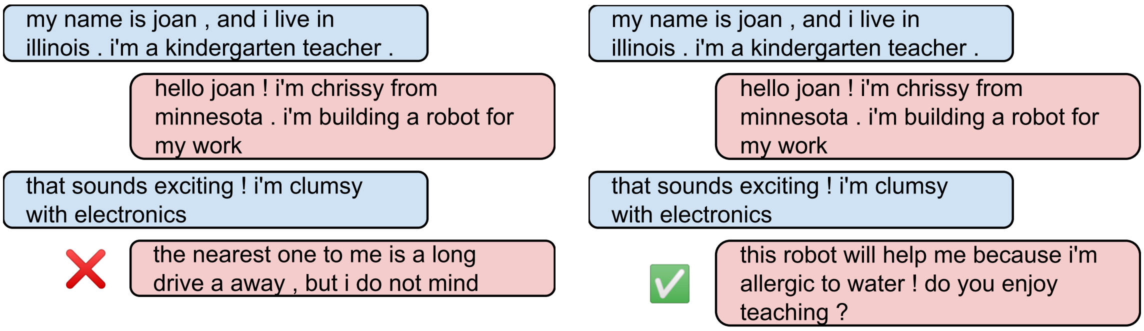
In this chapter, we make a general introduction of the work done by first presenting the motivation for the problems and challenges we want to solve. We then outline the key contributions of the work and finally explain the overall structure of the thesis.

1.1 Motivation

Computers have never been ever so present as they are today. It has even become possible to interact with them using natural language, e.g., voice assistants like Apple Siri and Amazon Alexa, or the chatbots on social platforms that let us solve tasks like booking a table or ordering dinner. While they are all beneficial for solving their jobs, most of us have experienced, on one occasion, problems in their operations. A typical scenario is how a user makes a specific inquiry, and the system's response is about something completely unrelated. As a relatively new technology, there is still a lot to be done regarding the research and development of such systems. One of the standard processes is evaluating the ability of a system to converse, i.e., the quality of response it creates in a dialogue. It is usually done by human annotators, making the development more resource-intensive and time-consuming. The research of dialogue systems can benefit from automated evaluation procedures.

In Figure 1.1, we present two dialogue conversations generated by dialogue systems participating in the ConvAI challenge [1]. It is easy to detect the difference between the two samples. The left one has a final response that has no relation to the history of the conversation. In contrast, the right one's last utterance is ideally within the context of the dialogue. The ConvAI challenge invited researchers and developers of dialogue systems to compete against each other. A winner was selected based on volunteers who had to converse with the dialogue system and give a score based on their experience. The competition could have quickly benefited from an automated metric that can quantify the capabilities of dialogue systems.

Revising the example dialogues in Figure 1.1, we see that they all use fluent language, which is good. Fluency is already one type of criteria that needs to be evaluated to understand the ability of a


my name is joan , and i live in illinois . i'm a kindergarten teacher .
hello joan ! i'm chrissy from minnesota . i'm building a robot for my work
that sounds exciting ! i'm clumsy with electronics
❌
the nearest one to me is a long drive a away , but i do not mind
my name is joan , and i live in illinois . i'm a kindergarten teacher .
hello joan ! i'm chrissy from minnesota . i'm building a robot for my work
that sounds exciting ! i'm clumsy with electronics
✅
this robot will help me because i'm allergic to water ! do you enjoy teaching ?

(a) Low-quality Dialogue

(b) High-quality Dialogue

Figure 1.1: Two sample conversations from the ConvAI challenge [1]. The example on the left demonstrates a response that does not follow the context and would leave a dialogue participant confused. On the right, there is a dialogue whose final response is coherent with the context.

dialogue system. However, this is arguably not the most difficult challenge for a modern dialogue system. Another more exciting and crucial criterion is coherency, i.e., how well does the dialogue system stay (or not) within context as seen in the samples. Unfortunately, here the possibilities are practically endless. Dialogue systems suffer from generic responses like: "I don't know," which can be considered a safe choice. However, such utterances do not benefit the overall flow of the conversation. Furthermore, if the response is awful, like in the example, it could lead to what is known as a "dialogue breakdown," a reply that breaks the flow of the conversation and makes it impossible to continue.

To measure the criteria mentioned above and detect dialogue breakdowns, we employ various techniques from machine and deep learning, which have already proven effective in natural language processing. Given the said family of approaches, one might be tempted to consider using supervised learning to solve the challenge of dialogue evaluation. However, while there is research doing that, we find it also reasonably trivial. Hence, a significant focus of this thesis is to solve the problem of dialogue evaluation using approaches that do not require direct supervision - either unsupervised or distantly supervised.

Overall, the vision of this work is to take steps in progressing toward an automated Turing test, i.e., an algorithm that can evaluate conversation just like a human does without having explicit training or education. Such a method would support the research and development of dialogue systems and allow a monitoring tool that can enable a fallback scenario once a problem occurs.

1.2 Problem Statement and Challenges

In this section, we specify the problem definition for this thesis. We then break it down into challenges that we discuss in greater detail.

Research Problem Statement

How can we automatically evaluate conversations to provide insightful feedback enabling the benchmarking of dialogue systems and requiring no human supervision or references for functioning?

As mentioned earlier, we observe the following challenges that need to be solved for unfolding our core problem based on our motivation.

Challenge 1: Developing an algorithm that does not require any labeled data

As we mentioned in our motivation, manual annotation of dialogues is time-consuming. Hence, we need to develop methods that do not depend on such annotated data, and thus, they have to be unsupervised or, at least, distantly supervised. While having a supervised algorithm could help reduce the workload for the short term, we need a more long-lasting solution that does not depend on human effort.

Furthermore, annotated data is just sometimes not available. Thus, unsupervised dialogue evaluation approaches would pave the way for algorithms that can be developed to address the issue of dialogue evaluation in various languages.

Challenge 2: Automatic evaluation should not require a reference

The most common way to evaluate dialogue systems is by employing a reference. In other words, a sample conversation context is presented, and the algorithm needs to generate a response. It is then compared to a reference utterance to estimate whether the dialogue system functions appropriately.

First of all, methods for comparing two sentences are still an active area of research. While some significant advances are made, the research community is still not satisfied with the state-of-the-art methods. Secondly, even if sentence comparison were a "solved problem," we have the issue of reference. In many cases, given a dialogue context, the space of possible responses is limitless. Furthermore, the problem partially overlaps with Challenge 1 since the availability of references could depend on manual effort.

Hence, the dialogue evaluation algorithms need to function as human annotators. They should be able to estimate the conversation quality without a reference.

Challenge 3: The automated evaluation methods need to be informative

From a linguistic perspective, there is a set of various criteria that a dialogue needs to meet to be correct. A better understanding of features like coherency or fluency would also enable researchers to develop dialogue systems to focus on the "right problems."

Revisiting our example in Figure 1.1, we saw that both target responses are fluent. However, one of them is not following the topic, i.e., it is out of context. At the same time, the other one could be considered ideally on topic. Therefore, it is beneficial to know the disadvantages and advantages of dialogue systems.

1.3 Research Questions

Research Question 1

Can anomaly detection methods be used to infer the quality of a dialogue?

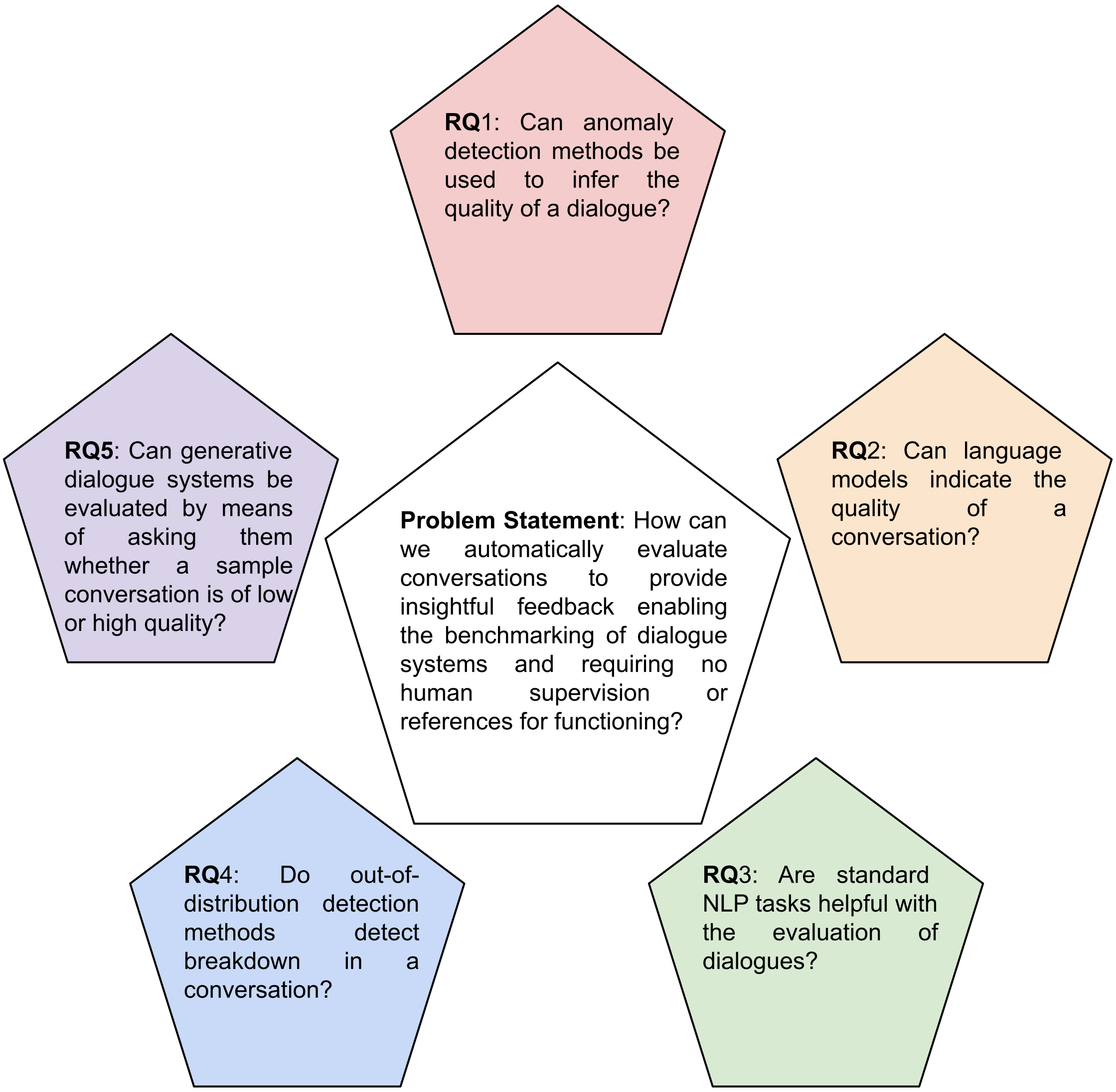

RQ1: Can anomaly detection methods be used to infer the quality of a dialogue?
RQ5: Can generative dialogue systems be evaluated by means of asking them whether a sample conversation is of low or high quality?
Problem Statement: How can we automatically evaluate conversations to provide insightful feedback enabling the benchmarking of dialogue systems and requiring no human supervision or references for functioning?
RQ2: Can language models indicate the quality of a conversation?
RQ4: Do out-of-distribution detection methods detect breakdown in a conversation?
RQ3: Are standard NLP tasks helpful with the evaluation of dialogues?

Figure 1.2: Overview of the Main Research Problem together with the five Research Questions.

There is the task domain of anomaly detection within machine learning, where systems need to detect samples that characterize themselves with distinctive features. However, some approaches do not need exemplary data on anomalies. This thesis hypothesizes that dialogue with lower quality or mistakes can be treated as anomalies. Hence, we apply standard anomaly detection methods from computer vision to evaluate dialogues. We assume that erroneous dialogues should appear anomalous, whereas correct ones - do not.

Research Question 2

Can language models indicate the quality of a conversation?

Language models have been part of NLP research and have found applications for many years.

Their main task is to learn to predict what is the next most probable word given a starting sequence. They have the advantage that they do not need any human-annotated data to be trained, but just raw text of any form, e.g., books or Wikipedia articles. In this thesis, we want to investigate if this property can be used to distinguish a low-quality conversation from a high-quality one. The assumption is that the former should receive higher probabilities from a language model while the latter has lower ones.

Research Question 3

Are standard NLP tasks helpful with the evaluation of dialogues?

Natural Language Processing deals with tasks that aim to replicate various language skills inherent to humans. For example, one such benchmark is semantic similarity, where NLP systems need to grade the semantic similarity between a pair of sentences, or the linguistic acceptability, in other words, fluency, of a piece of text. We hypothesize that such language understanding skills help indicate if a conversation is coherent or fluent. We investigate a standard set of such tasks that replicate various language skills and try to map them to different dialogue quality criteria. We expect these NLP benchmarks to correlate with multiple measures for dialogue quality.

Research Question 4

Do out-of-distribution detection methods detect breakdown in a conversation?

One of the problems that conversations, and consequently, dialogue systems, suffer from is a dialogue breakdown. A dialogue breakdown occurs whenever an utterance occurs such that it breaks the fluency and coherency of a discourse. In machine learning, a specific problem domain deals with the recognition of data samples that do not belong to a specified statistical distribution. Hence, they are different and stand out from the "normal" data points. This thesis hypothesizes that we can detect a dialogue breakdown if we take a similar perspective to utterances in a conversation. We assume that the "regular" discourse responses should be classified as belonging to the same distribution. In contrast, erroneous responses that lead to a dialogue breakdown should be perceived as out-of-distribution (OoD).

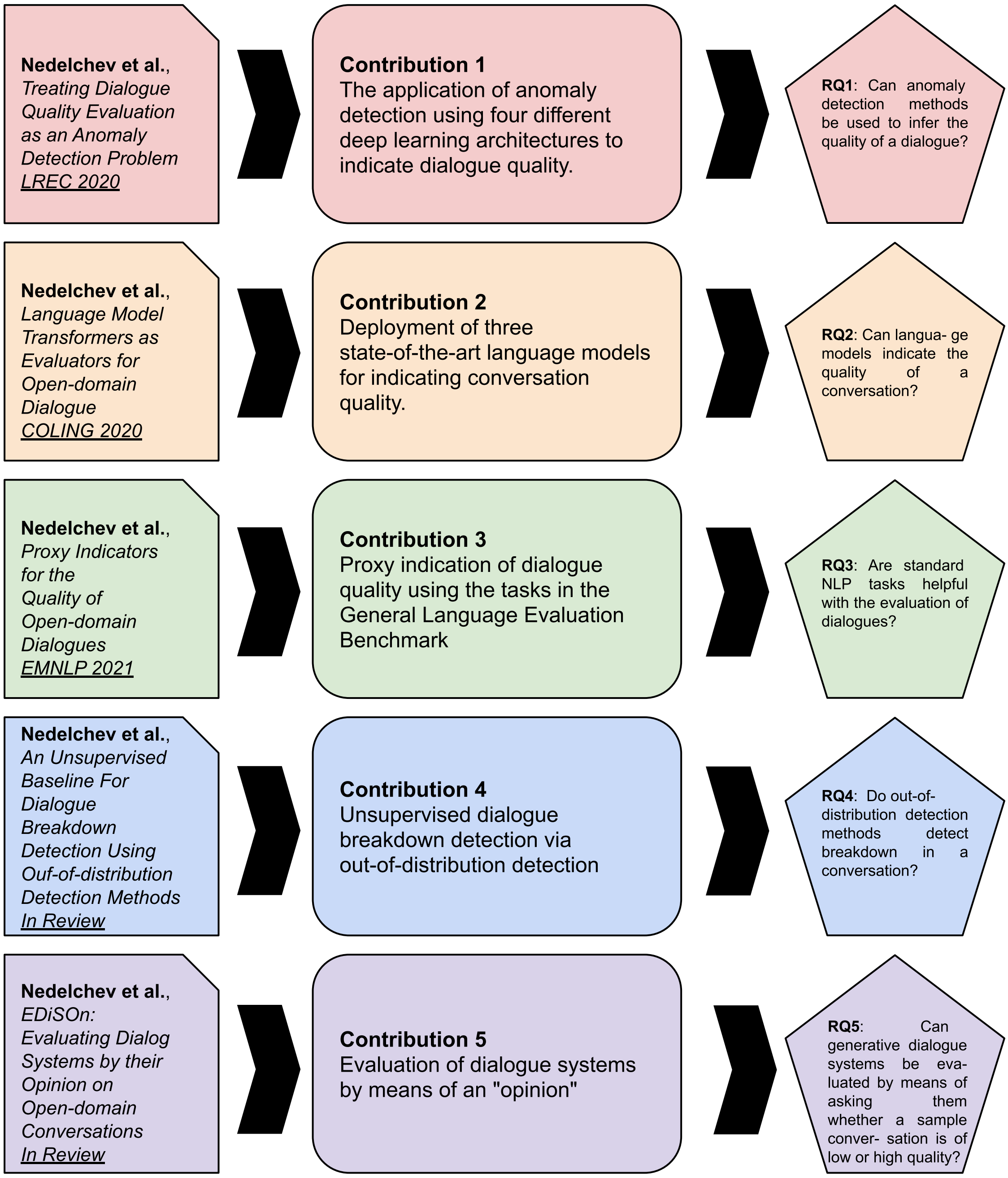
Research Question 5

Can generative dialogue systems be evaluated by means of asking them whether a sample conversation is of low or high quality?

Usually, dialogue systems are evaluated by means of a "reference solution." A system is provided with a sample dialogue history, and it has to generate a response, which is usually compared to a reference utterance based on the syntactic overlap. This approach is a problem for two reasons. First, it assumes that the response should use the exact words in the same order. However, the assumption does not necessarily need to hold for a dialogue system to generate a correct response. Second, the overlap in syntax also assumes an overlap in meaning. Yet, these assumptions do not need to hold for a correct utterance. For example, in most cases, whenever there is a question in conversations, there is more than one possible answer. Hence, we aim to evaluate dialogue systems by asking them for an opinion on whether annotated dialogue is of low or high quality. Furthermore, we assume that this opinion should correlate with human annotators.

1.4 Thesis Overview

In this chapter, we lay out the main contributions of the thesis. Furthermore, we present references to the scholarly articles that contribute to the research questions.



Nedelchev et al., *Treating Dialogue Quality Evaluation as an Anomaly Detection Problem* *LREC 2020*

Contribution 1 The application of anomaly detection using four different deep learning architectures to indicate dialogue quality.

RQ1: Can anomaly detection methods be used to infer the quality of a dialogue?

Nedelchev et al., *Language Model Transformers as Evaluators for Open-domain Dialogue* *COLING 2020*

Contribution 2 Deployment of three state-of-the-art language models for indicating conversation quality.

RQ2: Can langua- ge models indicate the quality of a conversation?

Nedelchev et al., *Proxy Indicators for the Quality of Open-domain Dialogues* *EMNLP 2021*

Contribution 3 Proxy indication of dialogue quality using the tasks in the General Language Evaluation Benchmark

RQ3: Are standard NLP tasks helpful with the evaluation of dialogues?

Nedelchev et al., *An Unsupervised Baseline For Dialogue Breakdown Detection Using Ouf-of-distribution Detection Methods* *In Review*

Contribution 4 Unsupervised dialogue breakdown detection via out-of-distribution detection

RQ4: Do out-of-distribution detection methods detect breakdown in a conversation?

Nedelchev et al., *EDiSOn: Evaluating Dialog Systems by their Opinion on Open-domain Conversations* *In Review*

Contribution 5 Evaluation of dialogue systems by means of an "opinion"

RQ5: Can generative dialogue systems be evaluated by means of asking them whether a sample conver- sation is of low or high quality?

Figure 1.3: Outline of the Contributions with regards to the Research Questions.

1.4.1 Contributions

Contributions to Research Question 1

The application of anomaly detection using four different deep learning architectures to indicate dialogue quality.

We implement anomaly detection for dialogue evaluation inspired by autoencoder neural networks. Intuitively, autoencoders detect anomalies by learning to create a "lossy compressed" representation of data points from which they can reconstruct the input. In the case of never-before-seen anomalous samples, their ability is strongly impaired. We demonstrate with four different architectures that the method has varying levels of moderate sensitivity to the quality of a conversation in a single overall score. Using anomaly detection is possible without any supervised training. Depending on the chosen architecture, we have slight variations as to which of those is more sensitive.

Anomaly detection for dialogue evaluation works by giving a score for an utterance of how anomalous it is. In other words, we measure how different it is from "ordinary" responses. Then, based on it, we derive an overall metric that can indicate the general quality of the target utterance, i.e., the higher the abnormality score, the lower the quality of the conversation.

Contributions to Research Question 2

Deployment of three state-of-the-art language models for indicating conversation quality.

Language models are a current trend in the field of natural language processing. They are proven to learn syntactic patterns and representations from regular pieces of text without using any additional labeled data. In the typical case, they are trained to predict the next word given a context. They predict a probability distribution over a vocabulary of possible words.

Using three prevalent language model approaches (BERT [2], GPT2 [3], XLNet [4]), we demonstrate that this ability can be used to indicate the overall quality of dialogues. Like humans, language models learn to "understand" text by "practicing to read it." Hence, we use the probability distribution over the words to derive a metric, which we calculate for each word in our distribution. We aggregate it on the whole utterance to tell us how fluent and coherent it is in one score.

Contributions to Research Question 3

Proxy indication of dialogue quality using the tasks in the General Language Evaluation Benchmark

For decades the computational linguistics community has been researching how to teach language skills to computers. Usually, it is done utilizing benchmarks that focus on various specific linguistic skills. For example, a benchmark would aim to train and test a system to detect similarities in meaning between pieces of text or classify the sentiment of a document (positive or negative).

The ability to participate in dialogue requires a mixture of those linguistic skills. This thesis demonstrates that these language abilities can be used as proxy indicators of dialogue quality. We use the General Language Understanding Evaluation (GLUE) benchmark [5], which contains diverse tasks focus on various aspects of language understanding. Furthermore, we demonstrate how to combine them into one score that estimates the overall dialogue quality. Finally, we explain also how we can control the "mixture" such that one can focus on specific conversation criteria more than others.

Contributions to Research Question 4

Unsupervised dialogue breakdown detection via out-of-distribution detection

One of the problems that dialogue systems often cause is dialogue breakdown. It is characterized by responses that are not related to the context of the conversation leaving the participants confused and unable to continue. The Dialogue Breakdown Detectching Challenge (DBDC) series [6, 7] aims to drive research and development to solve precisely this problem.

To the best of our knowledge, we propose the first unsupervised approach that can detect with reliable accuracy dialogue responses that could lead to a breakdown. We achieve this by standing on the shoulders of out-of-distribution detection methods, which can be seen as related as relatives of anomaly detection. Using a recent neural network architecture, DialoGPT [8], we demonstrate that OoD detection can be applied to identify utterances that can cause dialogue breakdown. Intuitively, we treat the problematic responses as unordinary.

Contributions to Research Question 5

Evaluation of dialogue systems by means of an "opinion"

The established procedure to evaluate dialogue systems is to look for overlaps with a reference response, which, as described, has some severe disadvantages. Hence, we demonstrate an alternative method for dialogue evaluation where we can "ask" dialogue systems for an opinion on whether a conversation is of low or high quality. This has the advantage that the dialogue cannot run into the problem of generating an alternative but equally good response since the possibilities are practically limitless.

To that end, the dialogue systems need to provide probabilities of each word in response. Then the set of scores is aggregated, and we apply a correlation analysis between it and a human annotator score. Ultimately, we use a set of architectures with varying complexity and progress to demonstrate that dialogue system approaches are comparable. The expectation is that for high-quality conversations, an approach has to generate high probabilities, and for low quality - the other way around. Indeed, the mode modern dialogue systems have higher correlations scores as expected.

1.4.2 Publications

The scholarly articles listed below serve as the scientific foundation of the thesis:

1. Conference Papers (peer-reviewed)

   a) **Rostislav Nedelchev**, Ricardo Usbeck, and Jens Lehmann. 2020. Treating Dialogue Quality Evaluation as an Anomaly Detection Problem. In Proceedings of the 12th Language Resources and Evaluation Conference (**LREC**), pages 508–512, Marseille, France. European Language Resources Association.

   b) **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2020. Language Model Transformers as Evaluators for Open-domain Dialogues. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6797–6808, Barcelona, Spain (Online). International Committee on Computational Linguistics (**COLING**).

c) **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2021. Proxy Indicators for the Quality of Open-domain Dialogues. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (**EMNLP**), pages 7834–7855, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

2. Papers in Review

a) **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2022. An Unsupervised Baseline For Dialogue Breakdown Detection Using Ouf-of-distribution Detection Methods. In Review for the 26th International Conference on Artificial Intelligence and Statistics (**AISTATS**).

b) **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2022. EDiSOn: Evaluating Dialog Systems by their Opinion on Open-domain Conversations. In Review for the 26th International Conference on Artificial Intelligence and Statistics (**AISTATS**).

1.5 Thesis Structure

The thesis consists of ten chapters. The current Chapter, Chapter 1 layouts the primary research question, the motivation for conducting scientific work, research questions, the respective contributions, and a list of publications that formally present the contributions.

Chapter 2 introduces core concepts and background knowledge that function as a basis for this work and support understanding the line of reasoning necessary for the thesis.

Next, Chapter 3 discusses efforts related to the core research question of evaluating dialogue systems. We review work done in reference-based and reference-free dialogue evaluation. In addition, we study state-of-the-art efforts in dialogue breakdown detection.

Chapter 4 studies anomaly detection for dialogue evaluation, where we experiment with four different architectures for dialogue modeling.

In Chapter 5, we report on how to use the language modeling approach for dialogue evaluation. We experimented with three LM methods - BERT, GPT2, and XLNet. In addition, we investigate the ability of some of them to act as dialogue systems.

After that, Chapter 6 investigates the usage of standard NLP benchmarks for conversation assessment. We perform experiments with the General Language Understanding Evaluation (GLUE) benchmark and BERT. Furthermore, we research the composition of different metrics into one.

Chapter 7 looks into dialogue breakdown detection. We research an unsupervised method inspired by out-of-distribution detection. We investigate three OoD detection approaches.

Second, to last, in Chapter 8, we work on an alternative method for evaluating dialogue systems. The core idea is to ask dialogue systems for an "opinion" rather than ask them to generate a possible "solution" to a problem.

Finally, Chapter 9 concludes the thesis by first reviewing the research questions and the contributions. In addition, it presents directions for future work.

CHAPTER 2

Preliminaries

> Consider your origin; you were not born to live like brutes, but to follow virtue and knowledge.
>
> *Dante Alighieri*

In this chapter, we present basic concepts that serve as the foundation for the work done in this thesis. First, we introduce deep neural networks and their specialized variations - Recurrent Neural Networks, Sequence-to-Sequence (Seq2Seq) Models, and Transformers, which are all critical milestones in NLP. After that, we discuss Natural Language Generation, and Dialogue Modeling, since they are an object of this work. Next, we present the state-of-the-art family of techniques, transformer-based language models (LMs). Finally, we conclude by visiting anomaly and out-of-distribution (OoD) detection.

We discuss all of the works and methods here on a rather superficial level without going into deep technical details. Hence, we kindly ask the reader to visit respective citation as for more information, as we go along with the preliminaries.

2.1 Deep Learning Fundamentals

Deep learning is currently one of the most researched sub-fields of machine learning. It uses artificial neural networks with many "layers." Hence, they have the name - "deep." This section follows the evolution of deep learning by discussing architectures that span from the first to the state-of-the-art ones.

2.1.1 Artificial Neural networks

A deep artificial neural network usually consists of multiple "layers." They can be split into three groups - "input," "hidden," and "output" layer. Commonly, there are only one each for the input and output layer. The second category is called "hidden" because it is between the other two, and one does not directly interact with it. Most commonly, neural networks are compared to the structure of brain. They consist of "neurons", that interlinked with each other in a complex graph structure, which also has a direction of the information flow. Due to this directed flow, they are also called feed-forward neural networks. For example, in Figure 2.1, we show a sample deep neural network as a graph that has a total of four layers - one input, two hidden, and one output layer. This type of network is often called also a feed-forward network, because there is a straight flow of information from the input to the output layer.

On a more local scale, one can perceive each of the nodes in the hidden layers of the neural network as neurons. Looking at the big picture, a deep neural network contains nodes that are characterized by a vast amount of learnable weights.

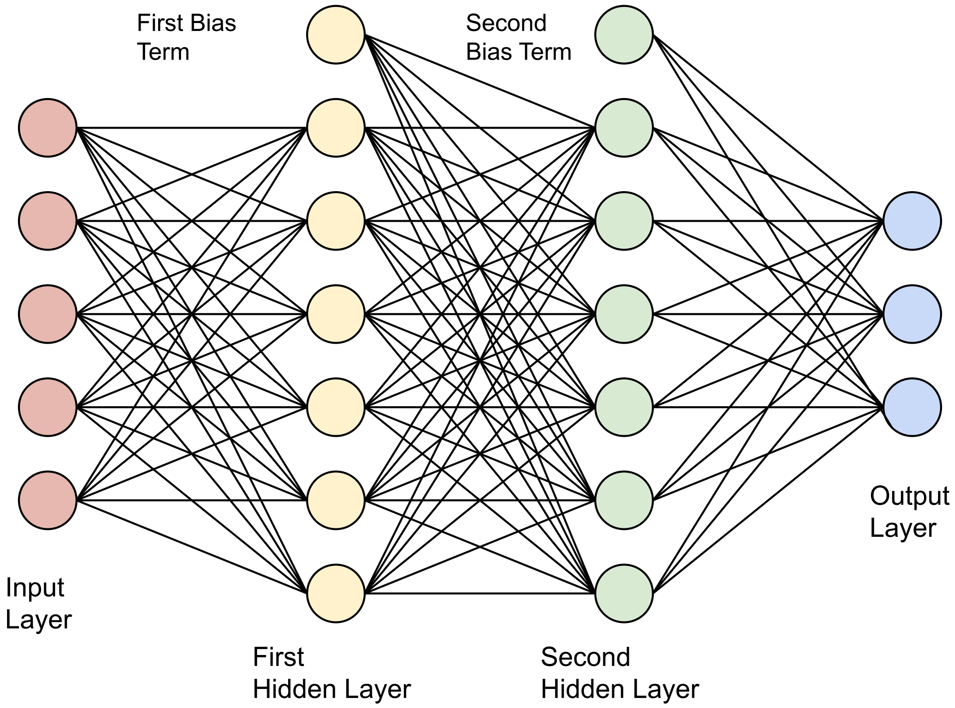


Figure 2.1: An example for an artificial neural network.

Artificial neural networks employ activation functions for each of the neurons. Their purpose is to decide whether a neuron should be "activated" and what degree. While the activation functions are an active research field, there are already a few established activation functions - sigmoid, softmax, Rectified Linear Unit (ReLU), Gaussian Error Linear Unit (GELU) [9].

Neural networks need to learn their characteristic weights from data. The first important step that enabled that is backpropagation [10]. Furthermore mathematical optimization helps with the training of ANNs. Formally, the problem is defined using the transformation $y = f(x)$ and an error (also commonly known as a loss) function that measures the difference based on obtaining a value for the dependent variable, $\hat{y}$, and some reference pairs of x and y.

Due to the sheer amount of learnable parameters that neural networks have, they employ a family of optimization procedures known as Stochastic Gradient Decent (SGD). It requires a (preferably massive) set of references pair of the independent and dependent variable, (x, y). Then, the transformation is applied to a sample of x, during which gradients are collected for each of the "neurons." Once the calculation is complete, an error is calculated, whose results are backpropagated from the output layers to the input one to adjust the learnable weights such that the error is minimized.

2.1.2 Recurrent Neural Networks (RNNs)

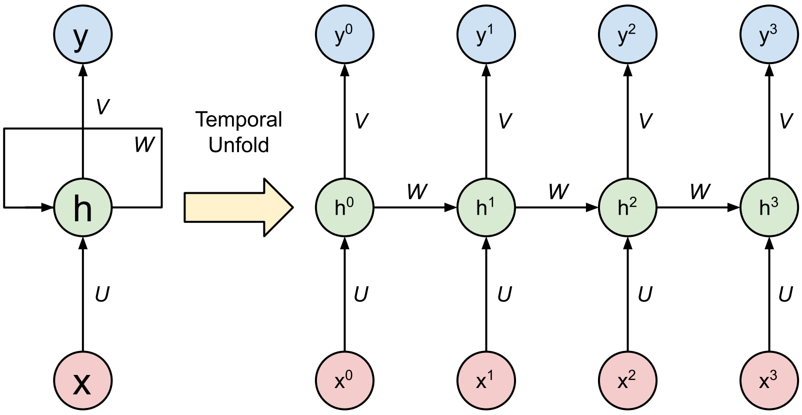


Figure 2.2: A demonstration of a recurrent neural network.

The research community focused on drafting new architecture for neural networks to accommodate specific properties for certain tasks. In contrast to feed-forward neural networks, Recurrent Neural Networks focus on modeling temporal sequences, i.e., capturing the information that happens during the transition between the sequence items. Furthermore, RNNs do not require that their input has a fixed size. Hence, they find widespread usage in natural language processing, where spoken or written language can be seen as a dynamic temporal sequence. In Figure 2.2, we show how the recurrent neural networks is time-wise unfolded, where *U*, *W*, and *V* are weights mapping the data from input, to hidden, and then to output neuron.

The main advantage of recurrent neural networks is that, theoretically, they can model long-term dependencies between states. However, it has been discovered that the bigger the distance between two states is, the more difficult it is for RNNs to capture that information [11]. As a result, they suffer from memory loss.

Thankfully, Long Short Term Memory (LSTM) networks, a specialized type of RNN, have been proposed to handle this issue [11]. Its core idea is to use an additional vector parameter, cell state, which acts as a pipeline that lets information flow through the whole chain of temporal states. However, LSTMs need to be selective to what flows through since a vector could store only a limited amount of information. Therefore, it adds neural "gates" that control how much the LSTM should remember or forget. There is a gate that decides how much to forget from the previous state, and another one one administers how to include from the current state. Finally, the third gate decides how much to output.

Gated Reccurent Unit (GRU) aims to supersede LSTM by making some efficiency improvements [12]. Namely, the input and forget gates are merged into one that decides a trade off between how much is forgotten or used from the input.

2.1.3 Sequence-to-sequence Models

The rise of recurrent neural networks gave birth to the sequence-to-sequence (Seq2Seq) model architecture [13]. Its aim is simple - it allows for the creation of a unidirectional mapping between two sequences. The most obvious example is machine translation, where we want to map a piece of text in one language to a piece of text in a different language. In addition, Seq2Seq models found a widespread application in dialogue systems where a conversation context can be mapped to a target response.

The intuition of the architecture is to create a separation of concerns and enable the specialization of two RNNs. The first has the responsibility to create a fixed-length representation of the input. Hence, it is often called an *encoder*. The second recurrent neural network can then use the representation to generate a new variable-length sequence. Therefore, it is called a *decoder*.

From a formal perspective, sequence-to-sequence models optimize the two RNNs jointly by maximizing the conditional probability between the input, $x = x_1, ..., x_n$, and output $y = y_1, ..., y_m$ sequences using the weights (w) of the model:

$$\max_{w} logP(y|x) \tag{2.1}$$

Reusing the established conventions from Section 2.1.2 and Figure 2.2, we present Figure 2.3 visualizing the architecture of the sequence-to-sequence model. We would like to turn the reader's attention that h^3 is the same vector representation passed from the encoder to the decoder and that we have two sets of weights, respectively, W_{Enc} and W_{Dec}.

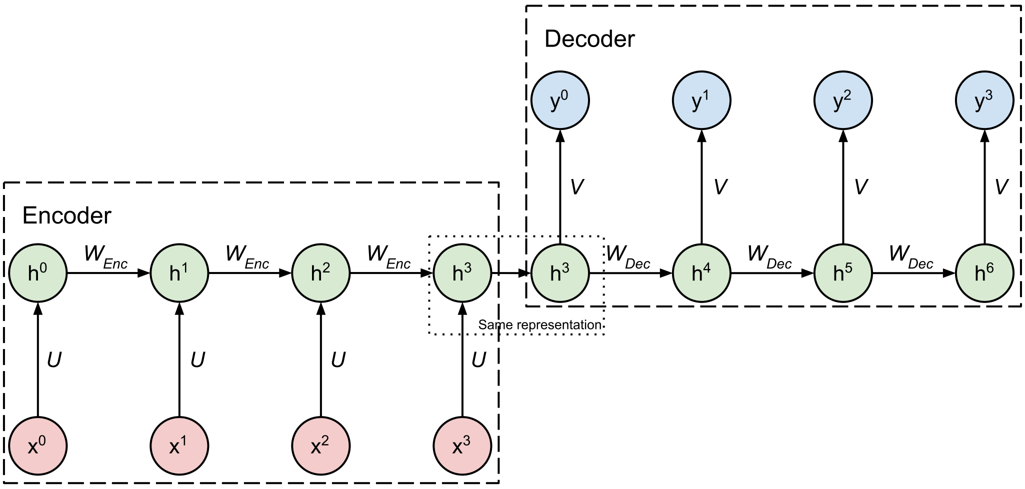
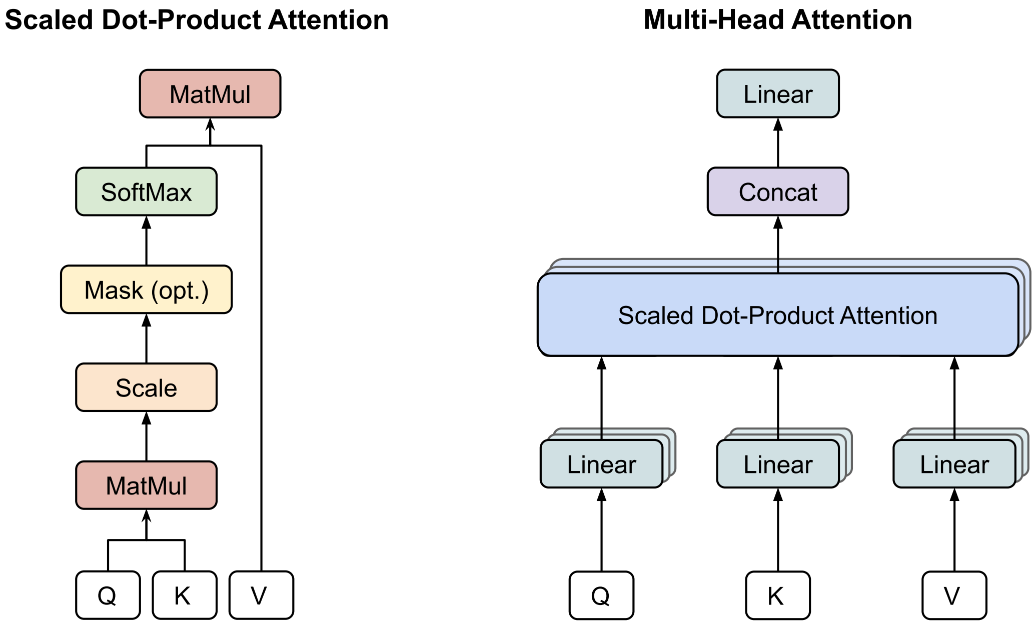


Figure 2.3: Architecture diagram of the sequence-to-sequence model.

However, the fixed representation passed from the encoder to the decoder acts as a bottleneck. Hence, Bahdanau [14] proposes the attention mechanism to alleviate the restriction and enable more information flow between the two RNNs. During each of its temporal steps, the decoder decides how much information to use from each input item. Hence, the encoder now has the single responsibility to model the dependencies within the sequence, but it does not need to create a fixed-length representation anymore. Instead, the decoder uses a single feed-forward layer and softmax to generate a distribution across the input token, deciding the degree of information flow.

2.1.4 Transformers

However, recurrent neural networks' most significant advantage turns out to be their greatest weakness, as well. Their sequential nature makes them rather slow to compute and does not allow parallel calculations. This is where the Transformer [15] steps in. It is a new architecture that replaces recurrence and counts on an attention mechanism to model dependencies globally on the whole sequence. Figure 2.4 visualizes the two core building blocks of the neural architecture.



(a) Scaled Dot-Product Attention

(b) Multi-Head Attention

Figure 2.4: On the left, Scaled Dot-Product Attention. On the right, Multi-head Attention with multiple parallel attention layers [15]

The inventors of the Transformers discuss the advantages of the new architecture. They claim that multiplicative attention (shown in Figure 2.4(a)) is the fastest and most efficient thanks to highly optimized implementations of matrix multiplication software. They further discuss that using multi-head attention (Figure 2.4(b)) enables the neural network to attend to the input information from different perspectives but also at other positions. Thanks to these features, the transformers can model global dependencies in the text without sequentially processing it.

2.2 Natural Language Generation (NLG) & Dialogue Modeling

One of the natural language processing (NLP) core fields is natural language generation (NLG). Initially and most notably, Reiter and Dale [16] describe NLG as a pipeline consisting of six distinct stages: 1. Content Determination, 2. Text Structuring, 3. Sentences Aggregation, 4. Lexicalization, 5. Referring Expression Generation, 6. Linguistic Realization. However, with the recent rise of neural networks, such staged approaches have become unnecessary. Deep learning is capable of learning representations that can model grammatical and semantic abstractions [17, 18].

There are two groups of neural network architectures suitable for dialogue systems. The first (causal) language models can predict the next word given a preceding sequence. For example, the work by Sutskever et al. [19] demonstrates the capabilities of LSTMs to predict the next character in a sequence. Hence, they are also commonly known as generative models due to their unparalleled ability to create language or other types of sequences like music

The second, encoder-decoder architecture [13] (often referred also as sequence-to-sequence or shortly, seq2seq), provides a decoupling between creating a fixed length representation of the input and consequently, decoding it into a sequence. There are many works that have used this approach to develop a dialogue systems [20–23].

Chen et al. [24] discuss a categorization based on their application - 1. task-oriented systems, 2. non-task-oriented systems. The former deal with assisting the user in completing certain tasks, e.g., booking a restaurant or finding out certain information. Task-oriented dialogue systems are usually pipeline-based, which means that they employ components with different concerns. The systems first comprehend the human message, represent it as an internal state, then perform actions in accordance with the dialogue state's policy, and lastly, turn the action into its surface form as natural language. Though not common, there are task-oriented systems that are also end-to-end but not as successful because there are domain specifics involved, and they are more difficult to incorporate.

The second category, which is also this work's primary target, is non-task-oriented dialogue systems. Unlike task-oriented systems, chatbots concentrate on open-domain conversations with humans. Given the nature of dialogue, sequence-to-sequence models find very widespread for creating non-task-oriented dialogue systems. However, these approach approaches have to meet some challenging criteria:

1. They need to provide context-sensitive responses, which requires the ability to model the complete history up to the current point.

2. The answers need to be diverse in nature. It is a well-known problem that dialogue systems use generic responses like "I don't know."

3. These approaches need to have awareness about the current topic and its "own personality" since these two often drive a conversation in a natural setting.

To converse in an open-domain setting, one needs to have standard knowledge. Without it, a dialogue system will not be able to chat. It requires an open-domain knowledge base. Finally, the "holy grail" of dialogue systems is to design an approach that can learn during an interaction with another human or another system, i.e., itself.

Knowing all of this, it is easy to imagine why open-domain dialogue systems are so difficult to evaluate, even for humans. Most of the time, overlap-based metrics are used to assess dialogue systems automatically. However, they are insufficient since they are unaware of the full range of valid responses for a conversation context. We revisit the topic in Chapter 3, Related Work, where we discuss previously done work.

2.3 Statistical Language Modeling

Since very recently, language models based on transformer neural networks [2–4, 15] are enjoying great popularity.

The first application of n-gram-based language models is recorded in the mid-1970s by two independent works of Jelinek [25] and Baker [26]. Given a sequence of tokens, $T = \{t_1, ..., t_N\}$, a forward language model computes the probability of the sequence by modeling the likelihood of a token t_K ($K \leq$ N), which has a history up to the K-th token [27]:

$$P(t_1, t_2, ..., t_n) = \prod_{k=1}^{N} P(t_k | t_1, t_2, ..., t_{k-1}) \tag{2.2}$$

Some of the initial neural network models [28] use initially a context-independent vector representation for a token, which all pass through one or more LSTM layers [11]. Then, they produce a context-dependent vector that serves as input to a softmax layer to predict the next token. In a reversed fashion, backward LM uses the context to the right of the target token to predict it. In contrast, bi-directional language models use a combination of both to predict the target word:

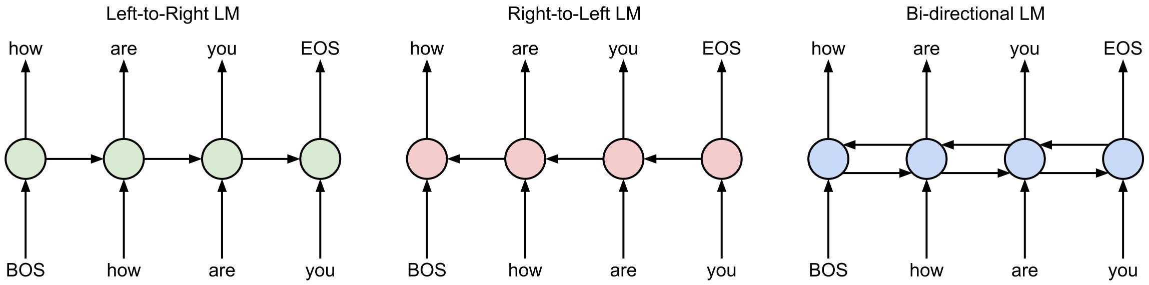


Figure 2.5: Diagram presenting the possible flows in language models. BOS stands for Begining of Sentence, while EOS - End of Sentence.

Radford et al. [3] propose generative pre-training (GPT2), where they use the transformer [15] as a forward (a.k.a. left-to-right) language model due to its superiority in terms of long-term memory when contrasted to recurrent neural networks like LSTMs.

Furthermore, Devlin et al. [2] suggest an innovative way to train language models, also utilizing transformers, specifically Bidirectional Encoder Representations from Transformers (BERT). They invent the masked language model (MLM), where a random subset of tokens from a sequence is masked or replaced, which the model then predicts using the remaining original context. Furthermore, BERT uses an additional LM objective: next sentence prediction (NSP). It works by teaching a model to recognize whether two sentences appear sequentially in a corpus or not.

Yet another innovative transformer-based language model is XLNet by Yang et al. [4]. It combines the best features of a generative LM like GPT2 and a masked LM like BERT by proposing to use the permutations of all factorization orders of a sequence to train. Thanks to it, XLNet learns to utilize knowledge from both sides of the target token and the respective context of other positions. Golovanov et al. [29] demonstrate that pre-trained transformer language models provide benefits for conversational agents.

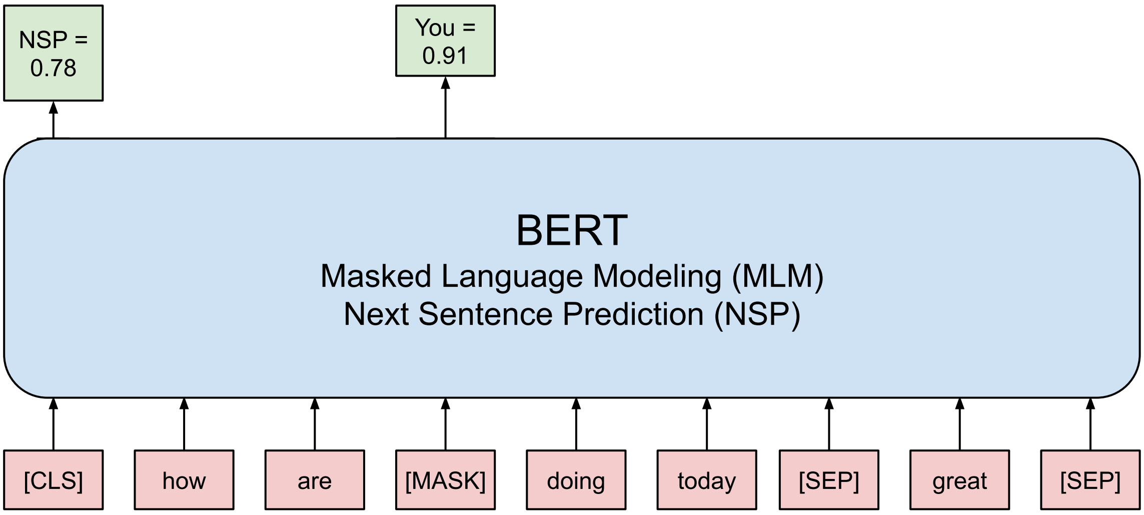


Figure 2.6: An example demonstrating BERT's language modeling objectives - Masked Language Modeling and Next Sentences Prediction. The probability for the masked hidden is predicted to find the most suitable candidate. In addition, BERT predicts the probability of the next sentence to model the cohesiveness of the two sentences [2]

For completeness, we mention other language models below that utilize transformers but are not integral to this work. We do not employ them in this work because the architectures discussed above already supersede them, or we deem their additions as not adequate for modeling dialogues.

Dai et al. [30] propose Transformer-XL, a new approach that allows transformers to model even longer sequences by caching and reusing intermediate hidden states. XLNet also utilizes the method in its implementation. Cross-lingual Language Model, by Lample and Conneau [31], introduces Translation Language Modeling, i.e., randomly masks words in parallel sequences in two languages to teach the model leveraging multi-lingual context. Liu et al. [32] present Robustly optimized BERT by just dropping BERT's next sentence prediction and a few other modifications in training. Raffel et al. [33] introduce the Text-to-Text Transfer Transformer, where the language-modeling objective is using a text-to-text perspective. Finally, conditional Transformer Language model, by Keskar et al. [34], incorporates conditioning on control codes to guide the generation of tokens.

Besides capturing syntax, LMs are also capable of modeling the semantics of sentences. The results of Tenney et al. [35] suggest that they can encode both syntax and semantics on a sub-sentence level. Furthermore, Zhou et al. [36] conducted a systematic benchmark to evaluate seven LM for their commonsense knowledge and reasoning. Their work suggests that they have a certain degree of those abilities. Commonsense is what would also help in evaluating open-domain dialogues.

2.4 Anomaly & Out-of-distribution (AD, OoD) Detection

Machine learning models have always operated under one core assumption. The data always comes from an independent and identically random distribution (i.i.d.). However, in reality, all variations of data occur frequently. Hence any ML approach can suffer from unseen data samples, including deep learning methods. In research, several domains focus on the recognition of such cases. Two of the fields are namely Anomaly Detection (AD) and Out-of-Distribution (OoD) Detection. While they have overlaps with each other, they also have some differences.

One of the first mentions of anomaly detection dates back to 1969 [37, 38], where it is defined as "samples that appear to deviate markedly from other members of the sample in which it occurs." This definition explicitly assumes a pattern that is followed by the majority of data points. However, to assert such a deviation, a distance metric needs to be used that is suitable for the target problem.

On the other hand, out-of-distribution detection deals with identifying test samples that are fundamentally different from the training data and, henceforth, should not be predicted into the known classes of the problem definition. It can be seen as a meta or addon to other tasks like multi-label classification or density estimation [39].

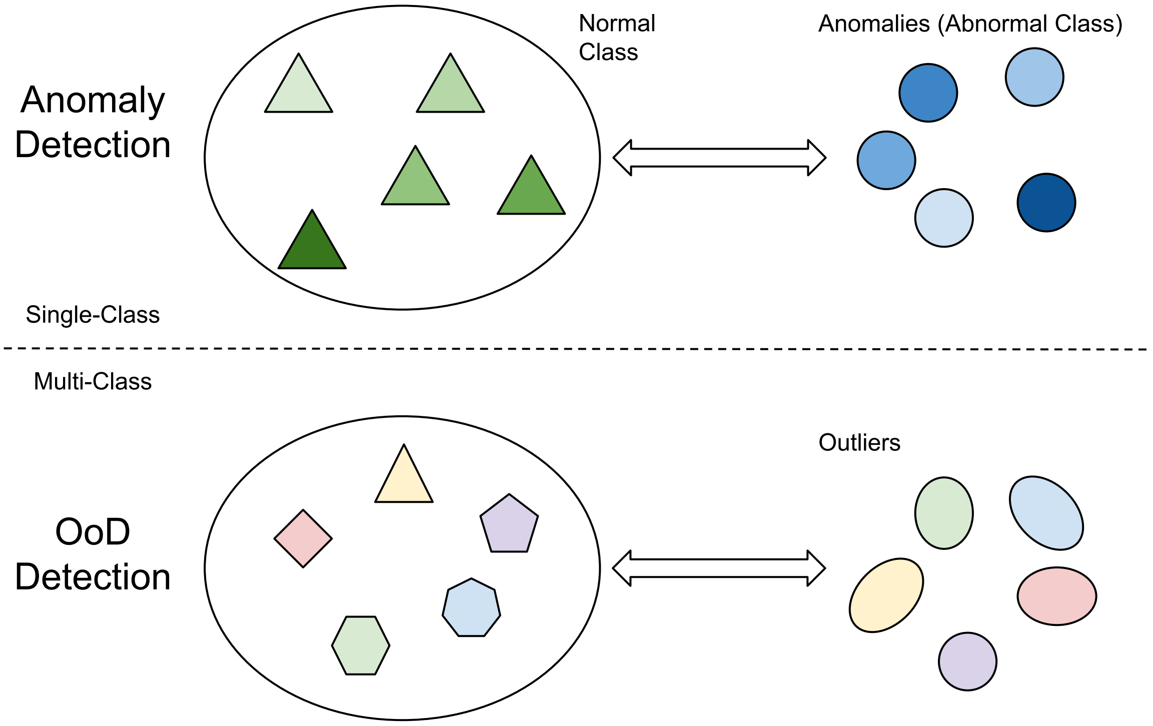


Figure 2.7: A visualization demonstrating the differences between anomaly detection and out-of-distribution detection.

For the purpose of this thesis, we revisit different works in anomaly and out-of-distribution detection. We review three OoD approaches that are pivotal to the thesis: **1.** Maximum probability of softmax-based classifiers [40]; **2.** Out-of-Distribution detector for Neural networks (ODIN) [41] and its generalized version [42]; **3.** Log-likelihood ratios [43].

2.4.1 Autoencoders

Autoencoders are a type of unsupervised neural network that learn how to create a compressed representation of the input data. They have two core components - encoder and decoder. The first one is responsible for creating the latent representation of the input data, which is usually of a lower

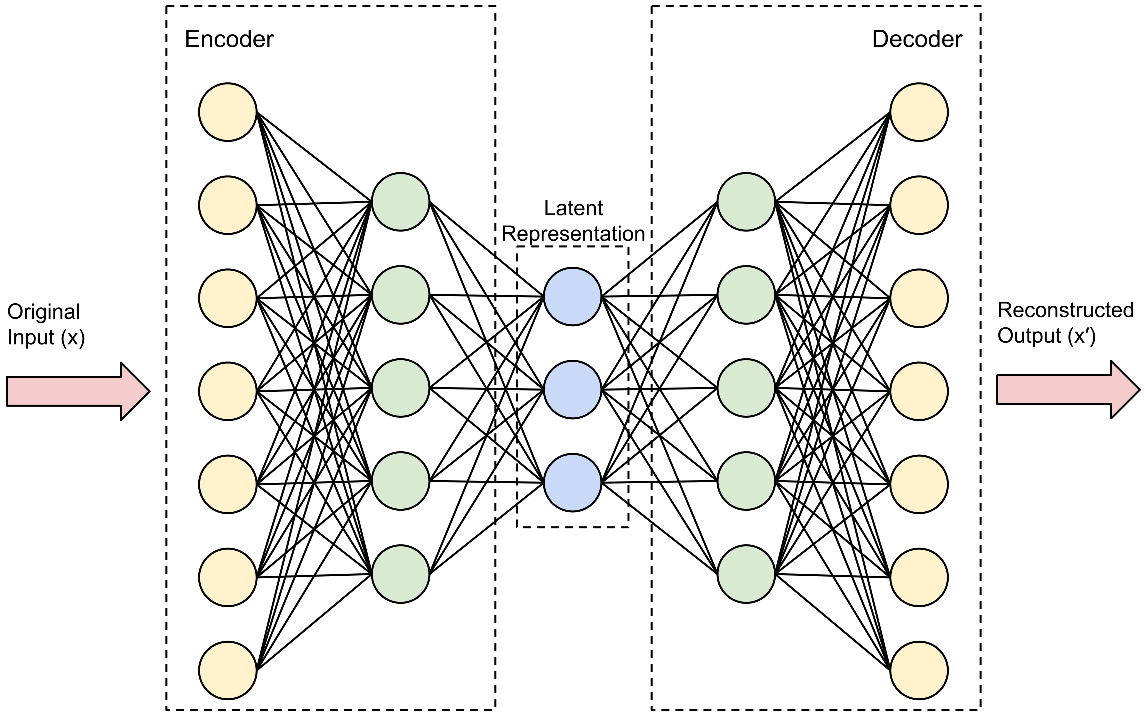


Figure 2.8: An example for an autoencoder neural network.

dimension. The decoder uses this compressed version of the data to reconstruct it to its original form. They are trained by optimizing a reconstruction error function that applies backpropagation through the whole autoencoder. It aims to minimize the difference between the initial input and its reconstructed version.

By using the aforementioned training regime, autoencoders learn key patterns and features of the input data. In the end, the reconstruction loss declines and falls to a minimum.

However, it has been established that they start to struggle whenever presented with a novel sample. By "novel," we mean a data point that does not follow the pattern of the data used for training. Hence, the reconstruction loss of these anomalous samples dramatically increases when passed through the autoencoder (shown in Figure 2.8), especially when compared to the in-train examples.

2.4.2 OoD Detection

The goal of OOD detection is to find test-times samples that are semantically distinct from the training data categories and so should not be forecasted into recognized classes [39]. For example, we consider a classification problem with ten classes. Focusing only a subset of those categories would render the remainder out of distribution.

We start with the work that first coined out-of-distribution detection as a term. Hendrycks and Gimpel [40] discuss the probability distribution of a softmax classifier and how it can be used for the task. In their work, they revisit various tasks (vision, natural language processing, and speech recognition) in deep learning and re-purpose existing datasets to create a benchmark for OoD detection. They propose to use the maximum softmax probability as an indicator for whether a data point is in or out of distribution. They demonstrate that the maximum probability for a softmax distribution is significantly lower for OoD examples than regular ones. Hence, this discrepancy has established the approach as a baseline for OoD detection approaches.

Liang et al. [41] propose ODIN (Out-of-Distribution detector for Neural networks). At its core, the method consists of two components. The first one is temperature scaling that is applied on softmax output:

$$S_i(x;T) = \frac{exp(f_i(x)/T)}{\sum_{j=1}^{N} exp(f_i(x)/T)}, \tag{2.3}$$

for the i-th output class, x is the input, and T is the temperature scaling coefficient. The second component is the application of random perturbations to the input:

$$\widetilde{x} = x - \epsilon \cdot sign(-\nabla_x \log S(x;T)), \tag{2.4}$$

where ϵ is perturbation magnitude. The authors report that this further pre-processing increases the softmax score gap between the in- and out-of-distribution data.

However, Hsu et al. [42] criticize the fact that both T and ϵ require OoD data to be tuned, which in certain use cases is not available. Hence, they discuss a decomposed confidence that consists of the joint class-domain probability and the domain probability:

$$P(y|d_{in}, x) = \frac{P(y, d_{in}|x)}{P(d_{in}|x)}, \tag{2.5}$$

x is the input, and y is the output class. To model those probabilities, they propose the following implementation for training:

$$P(y|d_{in}, x) = f_i(x); \tag{2.6}$$

$$f_i(x) = \frac{h_i(x)}{g(x)}; \tag{2.7}$$

$$g(x) = \sigma(w_g f^p(x) + b_g); \tag{2.8}$$

$$h_i(x) = w_i^T f^p(x) + b_i, \tag{2.9}$$

where $f_i(x)$ is the logit for the i-th class, $f^p(x)$ is the output of the penultimate layer of the network after applying the input, x, w and b are trainable parameters. For performing out-of-distribution detection inference, they suggest using $S_{DeConf} = max_i\ h_i\ or\ g(x)$. We report results using both.

In their work on image and genome sequence classification, Ren et al. [43] follow a similar intuition to Hsu et al. [42], where the OoD detector makes use of two components: a background component and a semantic component. The former models the population as a whole, whereas the latter captures patterns related to the domain data.

The background model is trained on perturbed in-domain data. Ren et al., [43] report using an independent and identical Bernouilli distribution with a rate of μ to decide which characters to be replaced with a random one. They report that $\mu \in [0.1, 0.2]$ achieves good performance for most of their experiments.

The log-likelihood ratio (LLR), i.e., the out-of-distribution detection score, is computed by using the probability scores from the background and semantic model:

$$LLR(x) = \log \frac{P_{\theta}(x_n|x_{<n})}{P_{\theta_0}(x_n|x_{<n})}, \tag{2.10}$$

where P_{θ} and P_{θ_0} are the softmax probabilities from semantic and background models, respectively. x_n is the n-th token, preceeded by the $x_{<n}$ tokens. x represents the concatenations of context utterances and response.

CHAPTER 3

Related Work

> Everything flows and nothing abides; everything gives way and nothing stays fixed.
>
> *Heraclitus*

This chapter reviews the community efforts related to the main research questions and the challenges outlined. First, we discuss approaches for dialogue evaluation using a reference since the domain acts as a starting point for the research goal of the thesis. After that, a section that revisits reference-free approaches follows, where works with similar motivation are presented. Next, we discuss previous works from the Dialogue Breakdown Detection Challenge series (DBDC) [6, 7] as a separate section since it is a well defined NLP benchmark. Next, we discuss the General Language Understanding Evaluation (GLUE) Benchmark since it serves as foundation for one one of our approaches.

3.1 Reference-based Dialogue Evaluation

Significant works in text summarization and machine translation have already proposed their field-specific metrics for automated assessment. For the former Recall-Oriented Understudy for Gisting Evaluation, ROUGE [44] is the most popular set of metrics for the problem. First is ROUGE-N, which measures n-gram (contiguous sequence of *n* items from a given text sample) overlap between system-generated response and reference. The most common versions of the metric are with uni- ($n = 1$) and bi-grams ($n = 2$):

Next is ROUGE-L, which uses the Longest Common Subsequence (LCS) instead of n-grams. The idea here is that a longer shared sequence would indicate more similarity between the two sequences. In addition, there are the less popular ROUGE-W, ROUGE-S, and ROUGE-SU, which introduce the usage of weighting mechanisms and skip-grams.

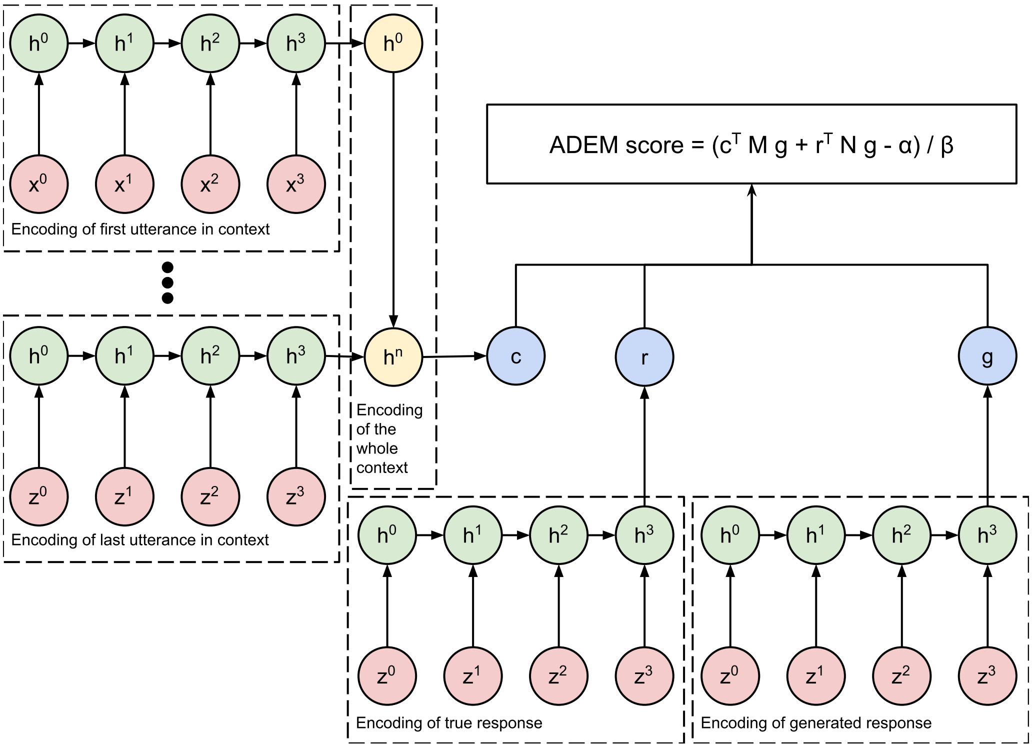
We move to two machine-translation-focused metrics. First is Bilingual Evaluation Understudy (BLEU) [45]. The method works by counting n-grams in the candidate translation that match n-grams in the reference text, where a 1-gram or unigram represents each token, and a bigram comparison represents each word pair. Regardless of the word order, the comparison is made. The counting of matching n-grams has been changed to guarantee that the number of times the words appear in the reference text is considered, rather than rewarding a proposed translation that creates a large number of plausible terms.

Finally, we review the Metric for Evaluation of Translation with Explicit ORdering (METEOR) [46]. The difficulty with BLEU is that individual sentence scores suffer because the BP value is based on mean lengths throughout the whole corpus. To overcome this problem, METEOR replaces the accuracy and recall calculations with a weighted F-score based on mapping unigrams and a penalty function for wrong word order.

Dialogue system research [21, 47, 48] constantly uses these metrics. However, Liu et al. [49] show that these metrics based on word-overlap between prediction and references are not reliable for evaluating the usefulness of dialogue systems. Hence, the field should use more sophisticated methods that consider the previous utterances of a conversation and their semantic meaning. This gave rise to a new line of research that focuses on new natural language evaluation metrics.

Zhang et al. [50] propose BERTscore, which computes a similarity score for each token in the candidate sentence and each token in the reference sentence, comparable to standard metrics. Although, instead of precise matches, they use contextual embeddings to calculate token similarity. Their work has better correlation scores with human judgments and outperforms existing metrics in terms of model selection.

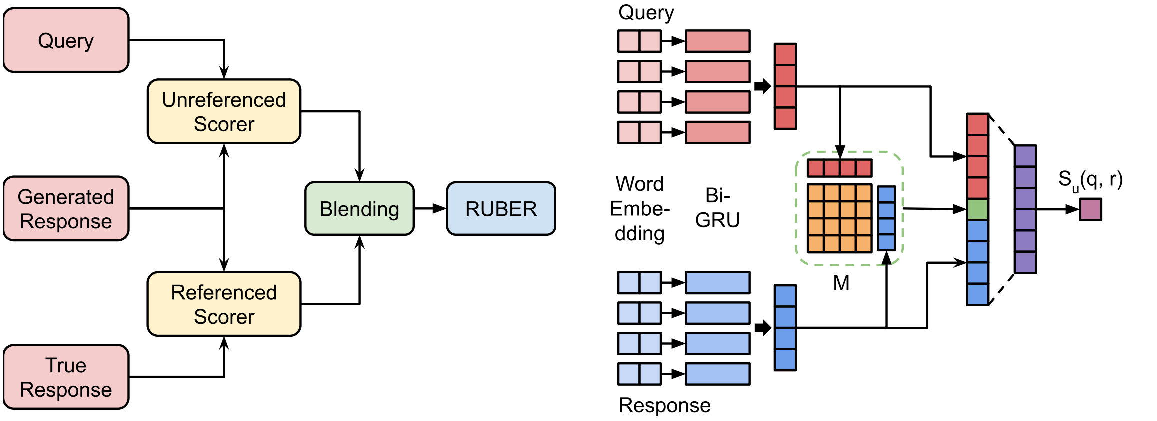
Lowe et al. present a cornerstone work in dialogue evaluation. [51]. They propose an *automatic dialogue evaluation model (ADEM)* (visualized in Figure 3.1) that employs a neural network approach that approximates human judgment using scored dialogues together with the context, reference response, and one generated by a dialogue system. Unfortunately, reference responses and human annotation scores are hard to obtain. It is challenging to employ the approach on large dialogue datasets.



h^0 h^1 h^2 h^3 h^0
x^0 x^1 x^2 x^3
Encoding of first utterance in context
ADEM score = (c^T M g + r^T N g - α) / β
h^0 h^1 h^2 h^3 h^n
c
r
g
z^0 z^1 z^2 z^3
Encoding of last utterance in context
Encoding of the whole context
h^0 h^1 h^2 h^3
z^0 z^1 z^2 z^3
Encoding of true response
h^0 h^1 h^2 h^3
z^0 z^1 z^2 z^3
Encoding of generated response

Figure 3.1: Architecture of the ADEM [51].

Another cornerstone is the work of Tao et al. [52], a *Referenced metric, and Unreferenced metric Blended Evaluation Routine (RUBER)* (show in Figure 3.2). They suggest a method consisting of two elements: The first one captures the resemblance between a generated and reference response using word vector pooling. The second one uses a neural network to estimate the relevance of a reply. The model is trained to distinguish whether an answer in a dialogue is the original or a random one from another conversation. A drawback of both approaches is that they use reference responses to derive a score.



(a) Overview of the RUBER approach

(b) Prediction of the unreferenced score using a neural network.

Figure 3.2: ARchitecture of the RUBER approach. On the left, we have the high-levevl

Furthermore, Sai et al. [53] demonstrate that machine learning approaches for dialogue evaluation like ADEM are susceptible to adversarial attacks. Even anything as simple as altering the word order in the text might cause ADEM to become confused. Experiments in various such hostile settings have produced unexpected conversation response ratings. They examine the scoring function suggested by ADEM in detail and tie it to linear system theory to foresee the system's flaws. Finally, they have devised an approach that can deceive such a system into giving a good rating to a response-generating system.

All of these approaches require a reference in order to evaluate a dialogue. It is a major disadvantage for these methods, since obtaining samples can be expensive. We saw that the early metrics inspired from other NLP tasks have downsides because they rely too much on the surface form of the text rather than its semantic meaning. On top of that, they are only consider one or best case, multiple possibilities, which as discussed do not cover the complete space of possible answers for a conversation history.

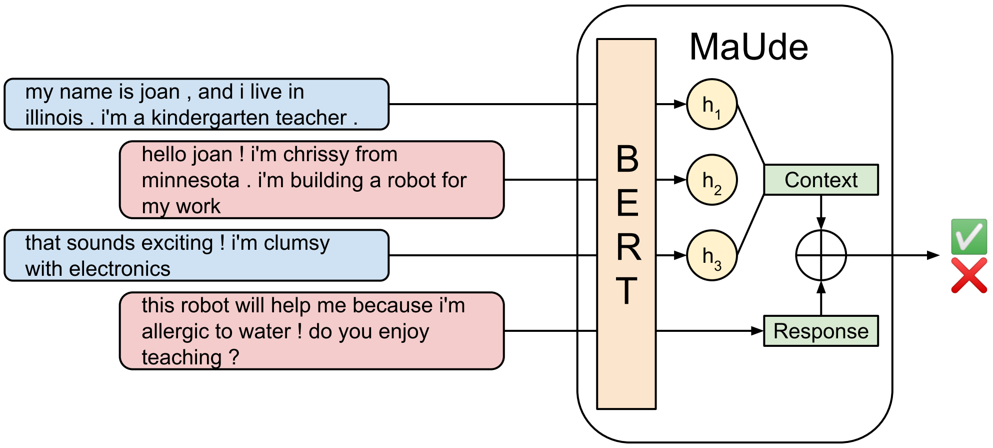
BERTScore [50], ADEM [51], and RUBER [52] could advance the reliance on semantic meaning. However, they still remain dependent on references.

3.2 Reference-free Dialogue Evaluation

While reference-based evaluation for dialogue is a sensible approach, it has a significant downside. The space of possible responses in a conversation is practically limited; hence comparing against one or even multiple references is impractical. Furthermore, obtaining these samples is resource intensive.

Hence, a big focus in the field is the research of approaches that do not require references to function, which is also the focus of this thesis.

Inspired by RUBER [52] Sinha et al. [54] propose a Metric for automatic Unreferenced dialogue evaluation (MaUde), which uses state-of-the-art pre-trained language models, paired with an advanced discourse aware language encoder and contrastive training technique. Their experiments demonstrate that MaUde (shown in Figure 3.3) has a strong correlation with human judgments.


my name is joan , and i live in illinois . i'm a kindergarten teacher .

hello joan ! i'm chrissy from minnesota . i'm building a robot for my work

that sounds exciting ! i'm clumsy with electronics

this robot will help me because i'm allergic to water ! do you enjoy teaching ?

Figure 3.3: Architecture of MaUde [54].

MaUde is designed to output a scalar score with values ranging between zero and one. It estimates how suitable a reply is to a given conversation history. They claim the task to be akin to Natural Language Inference (NLI), which we shortly discussed in Subsection 3.4.2. Following NLI, they approach the task by setting up encoders $f_e^{\theta}(c)$ and $f_e^{\theta}(r)$ to encode the conversation context and the answer, respectively. Then, a combination function, $f_{comb}(\ldots)$, is applied to the history and response representations, followed by a classification function, $f_{(}\ldots)$, finally concluded with a sigmoid, σ, function to normalize the score between zero and one:

$$\text{score}(c, r) = \sigma(f_t(f_{comb}(f_e^{\theta_1}(c), f_e^{\theta_2}(r)))) \tag{3.1}$$

Driven by the transformer-based language model trend, Sai et al. [55] present DEB (Dialog Evaluation using BERT). Inspired by BERT's next sentence prediction, they define the goal of predicting the future answer as determining if the provided response is a legitimate next response for the given context. The formal definition follows: provided with a dialogue context $C = \{w_1^c, \ldots, w_n^c\}$ and its response $R = \{w_1^r, \ldots, w_n^r\}$ through BERT and retrieve the representation , H_{CLS} , of the whole conversation. The final scoring is done by applying - $\hat{y} = \text{softmax}(WH_{CLS})$. In addition, they also perform the standard masked language modeling objective for training DEB.

To evaluate their approach, they propose an extension to the DailyDialog [56] dataset that includes five relevant and five adversarially crafted irrelevant replies for each conversation history. DEB demonstrated to have higher correlation coefficients than other reference-based approaches.

Also following the trend with language models, Mehri & Eskenazi [57] propose USR, an UnSupervised and Reference-free evaluation metric for dialog. Unlike DEB, it utilizes on RoBERTa [32].

They propose two components in their metric. In contrast to BERT, RoBERTa does not have a next sentence prediction training objective. Hence, as the first component of USR, they use the masked language training objective of RoBERTa. Given the dialogue context $C = \{w^c_1, ..., w^c_n\}$, its response $R = \{w^r_1, ..., w^r_n\}$, and their concatenation, they iteratively mask each word in R and compute its likelihood. The final metric of the component is computed as the sum of each probability. The MLM-based approach (illustrated in Figure 3.4) offers sort of a criterion that involves two aspects - fluency, i.e., how grammatical is the language, and common sense, since it is known that language models are capable of encoding some common knowledge.

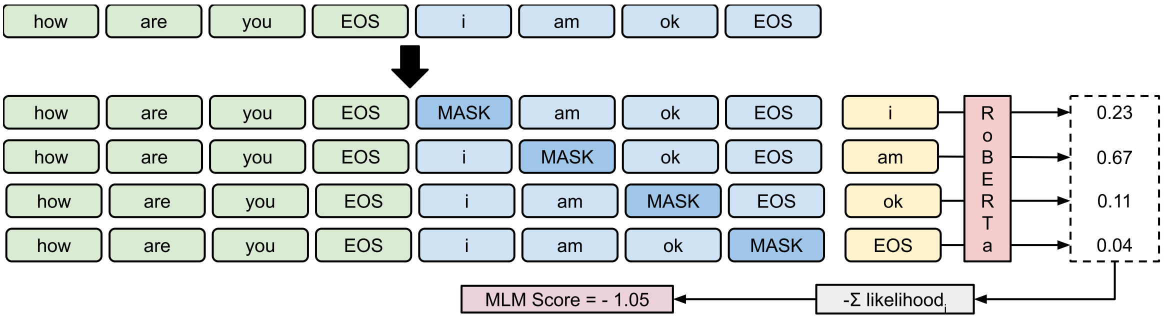

how are you EOS i am ok EOS
how are you EOS MASK am ok EOS
how are you EOS i MASK ok EOS
how are you EOS i am MASK EOS
how are you EOS i am ok MASK
i
am
ok
EOS
RoBERTa
0.23
0.67
0.11
0.04
MLM Score = - 1.05
-Σ likelihood$_i$

Figure 3.4: Architecture of the first MLM-based component of USR [57].

The second component is inspired by dialogue retrieval (DR). The same model from the first component is further fine-tuned for the retrieval goal. The format definition of dialogue retrieval is as follows: A dialogue context c, a response r, and a binary label y indicating whether r is the actual response or a random sample are used to train the model. On a high level, the architecture of the second component looks like the one of MaUde [54], where the significant difference is the usage of RoBERTa instead of BERT.

All of the approaches so far, DEB [55], USR [57], MaUde [54], utilize transformer-based language models. However, in all of the cases they use further training on dialogue dataset. Furthermore, the approaches provide only general quality scores without insights on the separate conversation criteria like fluency or coherency.

Gao et al. [58] propose DialogRPT. They utilize online forum feedback data (number of replies and upvotes) to construct a massive training set for feedback prediction. They transform the ranking issue into a comparison of answer pairs with a few confounding variables to reduce the possibility of a mismatch between feedback and interest. Based on 133M pairs of human feedback data, they trained their approach, a collection of DialoGPT-based models, and the resulting ranker exceeded various baselines.

They train their approach, in particular, the solve the following tasks:

- **updown** - How likely is the response to get the most upvotes on social media?
- **width** - How likely is the response to get the most direct responses from other users on social media?
- **depth** - How likely is the response to get the longest thread in terms of the number of chained responses by the other users?

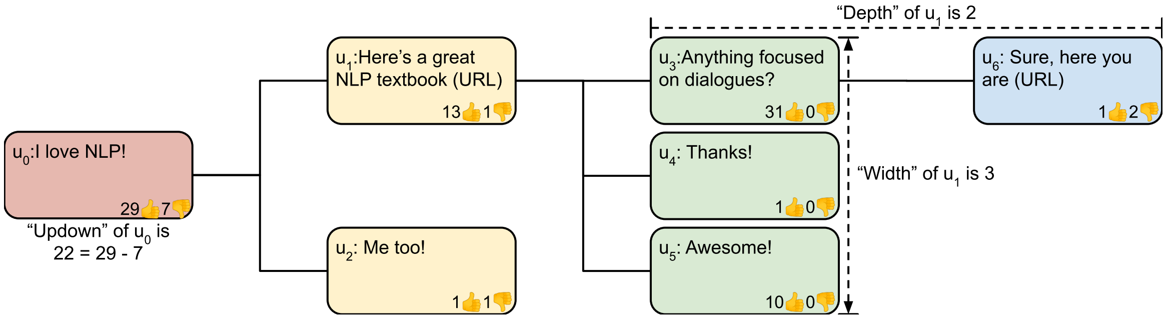


Figure 3.5: Modeling for online forum conversations as done in DialogRPT [58]

- **Human vs. Random** - How relevant is the response for the provided conversation context?;
- **Human vs. Machine** - How likely is it that a human rather than a machine has written the response?

The approach uses a contrastive learning objective together with the criteria mentioned above, where it learns to distinguish a "good" from a "bad" forum response.

Zhang et al. [59] propose a method that unifies turn and dialogue level evaluation - DynaEval (show in Figure 3.6). Their approach consists of four phases:

1. Derivation of contextualized representations of the utterances withing a regular, D, and a negative-sampled dialogue, $\bar{D}$, using SRoBERTa [60] and LSTMs.
2. Obtaining a directed dialogue graph, whose nodes represent the utterances and the edges - their temporal relations and respective speakers.
3. Inference of utterance representations based on the graph to model the interactions between the neighbors in the graph, i.e., the utterances.
4. Calculation of conversation-level metric indicates whether a negative-sample dialogue is preferred over a real one or the other way around.

While DynaEval does not provide scores for the various criteria, it delivers a comprehensive score that treats the dialogue as a whole by capturing all possible interactions.

Both works, DialogRPT [58] and DynaEval [59], present advanced approaches for feature engineering that enables the extraction of useful information from dialogues. The former uses dialogue criteria which are helpful but only indirectly infer the most fundamental ones like fluency and coherency. DynaEval proposes good techniques that model the whole dialogue and the interactions with it. However, the method proposes only an overall criteria rather than informing on specific ones.

3.3 Dialogue Breakdown Detection

In order for a user to receive meaningful replies through interactions with a chat-based dialogue system, it is crucial to ensure that communication is fluid. The majority of earlier dialogue research has not concentrated on preventing dialogue breakdown. One of the most significant obstacles is that

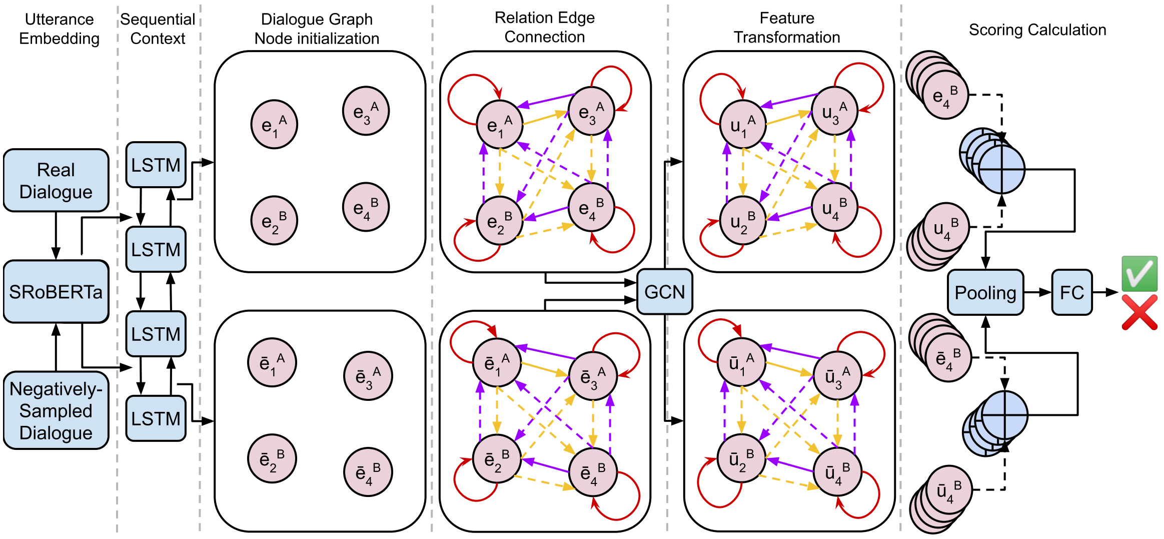


Figure 3.6: The construction of DynaEval. The input consists of two divergent conversations, D and $\bar{D}$. The result is a single score that indicates whether D is favored over $\bar{D}$. The utterance-level representation produced from the SRoBERTa model is utilized to initialize conversation graph nodes. Different arrows concerning edge connection denote various relationships: (1) Solid line denotes intra-speaker dependence. (2) Dotted line shows inter-speaker interdependence. (3) The color red signifies self-connection. (4) Purple color signifies a relationship between future statements to prior utterances. (5) Yellow indicates a relationship between past and future statements [59]

a conversation system may create an unintended statement, resulting in a dialogue breakdown that affects the quality of the interaction.

This problem has been targeted by a series of challenges, Dialogue Breakdown Detection Challenge (DBDC), to promote its research since it is a significant issue that requires attention. The first competition of the series was held in 2016 [61], and the most recent one with results is DBDC4 [7] from 2019. In this thesis, we focus on the newest one. The task is set up as follows. A conversation history consists of a sequence of alternating user and system utterances. The target utterance for dialogue breakdown detection is the next system utterance. Each instance is assigned one of three candidate classes: Breakdown (B), Possible Breakdown (PB), and Not a Breakdown (NB) by a group of human annotators. This setup offers a distribution over the possible labels since the evaluators never agree entirely. The output of a model includes two components: a predicted class from one of the three candidates B, PB, NB, which is compared against the majority vote of the annotators, and a probability distribution over the three classes. DBDC4 includes two tracks in two different languages (Japanese and English) with the same task setting. In this thesis, we work only with the English one.

Shin et al. [63] apply a neural network approach based on bidirectional LSTM. In addition, to compute the likelihood of each classification, the system uses global and local contextual information from human and system utterances. The embedding of each input (user speech, system utterance, or label) is delivered to a global-local attentive encoder (described below). Next, an attention module computes the external memory associated with system utterances up to turn (n-1) and user utterances up to turn (n) based on the encoder outputs. Finally, the external memory context, the query (current system utterance) context, the last user context, and the current label context are sent to the scoring module, which estimates the likelihood of the current label.

Sugiyama [62] proposes the usage of an ensemble of BERT models that aim at different dialogue-specific features in order to make a final prediction for breakdown detection. First of all, an instance of BERT is used to create a contextual representation of the conversation history. The embeddings of each utterance are average to deliver a fixed-length vector.

Furthermore, two more instances of BERT are used to estimate the dialogue act (types of an utterance such as question or greeting) of the dialogue context utterances and a prediction of what the next one would be based on the context so far. In addition, some hand-crafted features like sentence length or the number of utterances so far are also utilized. Finally, all of the described features are concatenated and passed through a feed-forward network to get the probability distribution over the three classes.

Wang et al. [64] use random forests combined with term frequencies and word embeddings to predict a probability distribution.

Co-attentive Cross-lingual Neural Model (CXM) by Lin et al. [65] utilizes the most recent advances in language models to tackle the DBDC4 challenge. They take advantage of a state-of-the-art cross-lingual pre-trained language model, XLM-R [66], which is pre-trained on large-scale multilingual corpora. Compared with other cross-lingual language models mBERT [2] and XLM [31], the data used to pre-train XLM-R is enlarged by orders of magnitude, especially for low-resource languages, which include both English and Japanese. In addition, they use a co-attentive encoder to compute a comprehensive representation of dialogue history and the target response to model their relationship better.

In this sub-section, we discussed multiple works that tackle dialogue breakdown detection using

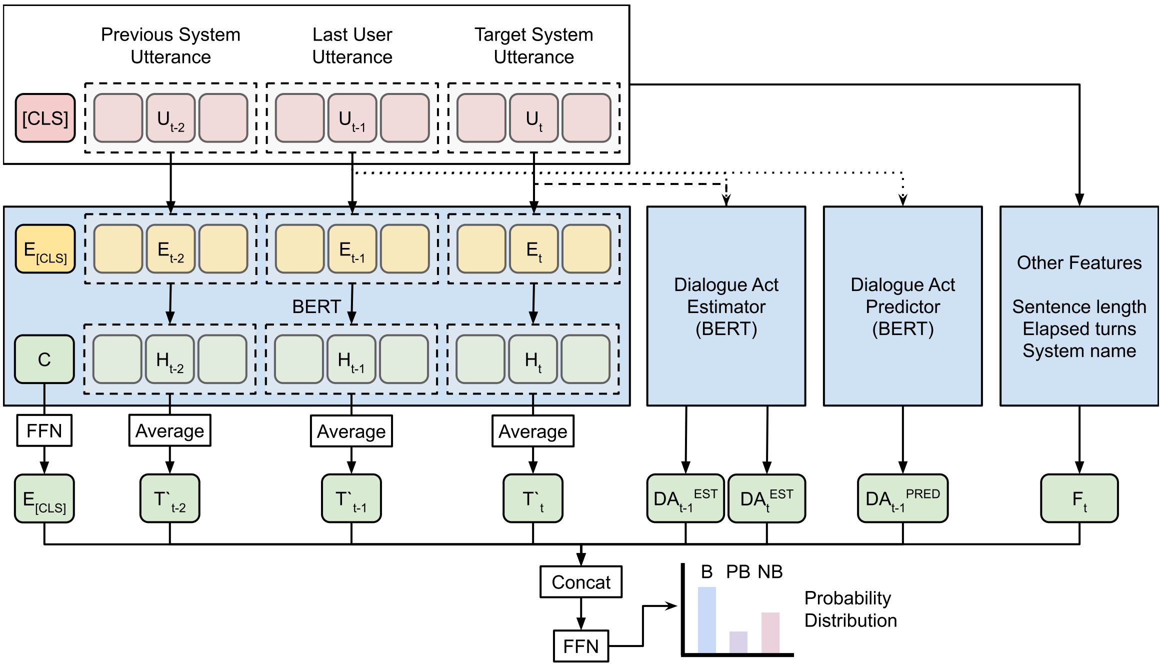


Figure 3.7: Architecture of BERT-based approach for dialogue breakdown detection combined with handcrafted features. [62]

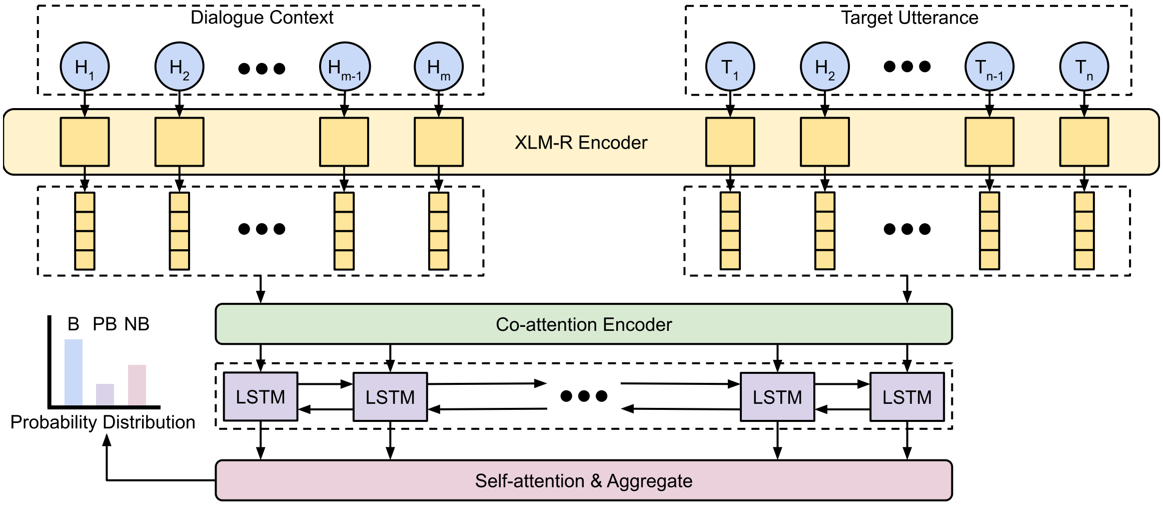


Figure 3.8: Architecture of Co-attentive Cross-lingual Neural Model [65].

approaches. Many of them apply advanced feature engineering and modern neural-network architectures. However, all of them are supervised and require training data in order to work. While the DBDC4 benchmark is already in two languages - English, and Japanese, this does not help with other languages. Hence unsupervised methods are necessary for target languages that are low on resources.

3.4 General Language Understanding Evaluation (GLUE) Benchmark

This section briefly introduces the General Language Understanding Evaluation benchmark [5], its sub-tasks, and their relevance to this work. GLUE has two categories of tasks - single- and pairwise-sentence tasks. They provide annotated data for training models to solve various natural language understanding problems. The section also discusses how these NLP tasks could be related to dialogue evaluation since they are initially irrelevant to this work's core topic. We present an overview of the whole benchmark in Table 3.1. The presentation of each of the tasks follows.

3.4.1 Single-Sentence Tasks

Corpus of Linguistic Acceptability (CoLA) [67] comprises samples in the English language that have scores for their grammatical correctness. Formally, this is a binary classification problem, where sentences are either acceptable (one) or unacceptable (zero) [5]. To evaluate dialogues, CoLA can provide fluency measures that show how grammatically sound a conversation is.

Stanford Sentiment Treebank (SST-2) [68] contains text excerpts from the movie reviews that have their sentiments annotated by humans as positive (one) or as negative (zero). Common sense would suggest that attitude provides no apparent relation to dialogue quality. Nonetheless, Ghandeharioun et al. [69] perform an ablation study as part of their work to see if knowledge distillation based on sentiment offers any benefits to evaluating a conversation. Their research shows that there can be an improvement depending on the neural network model and the target dataset. So, we investigate how it relates to annotator scoring on dialogue evaluation.

3.4.2 Pairwise-Sentence Tasks

The pairwise-sentence tasks consider a pair of utterances that appear sequentially in a dialogue.

Microsoft Research Paraphrase Corpus (MRPC) [70] is a dataset of sentence pairs extracted from news media, where each couple has scores as having the same meaning or not. Formally, it is a binary classification problem. A paraphrase has a label as positive, and non-semantic equivalence is negative. In the context of dialogues, a machine learning prediction for this task could imply that a response to an utterance is just repeating the former. At the same time, a partial degree could suggest some relevance. The negative case does not have a straightforward interpretation.

Quora Question Pairs (QQP) [1] is a corpus of question pairs extracted from the community question-answering platform Quora. Similar to MRPC, The focus is to flag a duo of questions as having the same semantics or not.

Semantic Textual Similarity Benchmark (STS-B) [71] is a dataset of paired-up media captions, news headlines, and sentences from natural language data that are given a similarity score from one to five by a human annotator. From a formal perspective, this is a regression problem where the output ranges between one and five. In a similar fashion to the last two tasks, this task can provide insights into the relevance and coherence of a response to its preceding utterance by assessing its semantic similarity.

Question Natural Language Inference (QNLI) [5] dataset is a re-adapted version of the Stanford Question Answering Dataset (SQuAD) [72]. The original dataset contains question-paragraph pairs, where an excerpt of the paragraph is an answer to the question. Wang et al. [5] convert it such that a question is paired up with each sentence from the context paragraph. Only the sentence with the answer to the questions has a label for textual entailment; the rest do not. The question is a hypothesis that could entail the sentence or not. It is treated as a relevance ranking problem, where a question can be more relevant to a sentence than others. Regarding dialogue quality, such a task can help with a response's relevancy assessment more straightforwardly than MRPC, QQP, and STS-B.

Recognizing Textual Entailment (RTE) datasets [5] consist of series of challenges: RTE1 [73], RTE2 [74], RTE3 [75], and RTE5 [76]. Pairs of sentences have been sampled from news and Wikipedia articles, which have been marked, similarly to QNLI, as textual entailment or no textual entailment[2], a binary classification problem. In a similar fashion to QNLI, RTE can be used to determine the relevancy of a response to an utterance. However, unlike QNLI, RTE does so for general statements rather than just questions.

[1] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

[2] Originally, there were two additional labels: neutral and contradiction. However, Wang et al. converted the two classes to no textual entailment [5].

Task	Description	Data Sample	Label	Metric
		Single-Sentence Tasks		
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one."	Unacceptable	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"rich veins of funny stuff in this movie"	Positive	Accuracy
		Pairwise-Sentence Tasks		
MRPC	Is sentence B a paraphrase of sentence A?	A) "The DVD-CCA then appealed to the state Supreme Court." B) "The DVD CCA appealed that decision to the U.S. Supreme Court ."	Not Equivalent	Accuracy F1
STS-B	How similar are sentences A and B?	A) "A man is playing the cello." B) "A man seated is playing the cello."	Very Similar	Pearson Spearman
QQP	Are the two questions duplicates?	A) "How do I lose weight fast?" B) "What is the best way to reduce weight?"	Duplicates	Accuracy F1
RTE	Does sentence A entail sentence B?	A) "Oil prices fall back as Yukos oil threat lifted " B) "Oil prices rise."	No Entailment	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What percentage of farmland grows wheat?" B) "More than 50% of this area is sown for wheat, 33% for barley and 7% for oats."	Answerable	Accuracy
MNLI	Does sentence A entail or contradict sentence B?	A) "I'll twist him, sir." B) "I'll make him straight."	Contradiction	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns. Is this the correct noun?	A) "I couldn't put the pot on the shelf because it was too tall." B) "The pot was too tall."	Correct Referent	Accuracy

Table 3.1: A table overview of all of the tasks in the General Language Evaluation benchmark. Data samples, together with the labels and evaluation metrics, are included.

Multi-Genre Natural Language Inference Corpus (MNLI) [77] is a compilation of sentence couples collected via crowd-sourcing that have been annotated for textual entailment, similarly to QNLI and RTE. However, MNLI does that as a three-class classification problem - textual entailment, contradiction, and neutrality. The task is not used for the work due to the lack of a straightforward mapping of those three classes to an ordinal/continuous variable like a dialogue quality score.

Winograd Schema Challenge (WNLI) [78] aims at reading comprehension where a system must gain an understanding of a sentence with a pronoun and then choose the suitable referent from a list of choices. Due to its nature, this task is not relevant and not used for this work.

CHAPTER 4

Anomaly Detection for Dialogue Evaluation

> The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'
>
> *Isaac Asimov*

Recently, machine-learning powered dialogue systems have been gathering much attention from industry and academia alike [24]. These systems have applications in various contexts, starting from personal speech assistants like Amazon Alexa or Apple Siri, through the "chatbots" on instant messaging platforms like Skype or Slack. Nowadays, researchers and developers who work on dialogue systems rely mostly on human annotators to evaluate the quality of a conversation [1, 48, 79]. This can be very costly in terms of resources. Thus, the research and development of these systems could benefit significantly from an automated approach that can evaluate conversations.

Human annotators distinguish low from high-quality dialogues similarly to anomaly detection. Conversations generated from computer systems can appear to human annotators as very unusual, i.e., an anomaly. Their perception is based on extensive conversational experience with real people, rather than using an explicit reference that helps to determine what is correct or wrong.

Research Question 1

Can anomaly detection methods be used to infer the quality of a dialogue?

Thus, the main contribution of this chapter is to investigate whether dialogue modeling approaches used for dialogue systems can detect anomalous conversations in contrast to normal ones. To the best of our knowledge, this is the first paper that attempts solving dialogue evaluation by treating it as an anomaly detection problem.

The chapter is based on the following article:

> **Rostislav Nedelchev**, Ricardo Usbeck, and Jens Lehmann. 2020. Treating Dialogue Quality Evaluation as an Anomaly Detection Problem. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 508–512, Marseille, France. European Language Resources Association.

In this chapter, we investigate four approaches for dialogue modeling that are considered major milestones in the development of dialogue systems. We adapt them to an anomaly detection task. The models provide anomaly detection prediction on utterance-level within a dialogue, which we use for a correlation analysis with human evaluators

4.1 Background

Larson et al. [80] propose outlier detection to detect erroneous utterances within a dialogue for clean data annotation in an NLP dataset. The approach averages word embedding of a reply's content to obtain an utterance level representation. After that, the second stage clusters the vectors, and the top-N[1] are considered anomalous. The approach provides no dialogue-level information about the coherency of the conversation and does not offer a replacement for human annotators. So far, this is the only known work that could be considered related to our specific problem of anomaly detection combined with dialogue systems to the best of our knowledge.

Anomaly detection, very commonly also outlier or novelty detection, deals with the problem of finding instances of data that do not belong to the regular pattern like most of the others [37].

There is a long list of works in NLP that have considered using anomaly detection for discovering incorrect annotations [80–82]. Most of them use handcrafted features to solve the problem.

In the field of deep learning, autoencoders found usage in significant amounts of research to solve problems from various domains. According to Chalapathy et al. [38], they are at the core of all unsupervised neural-network-based anomaly detection methods. They have found application in a wide variety of domains like intrusion or malware detection, bank, or insurance fraud. Autoencoders learn to create another representation of data (usually, one of lower dimension) and then reconstruct from it the original input. Their effectiveness is measured using a reconstruction error. Thus, on examples that an autoencoder has observed and trained on, it has a lower reconstruction rate. At the same time, on rare or not-previously seen samples, it will exhibit a consistently higher error.

4.2 Methodology

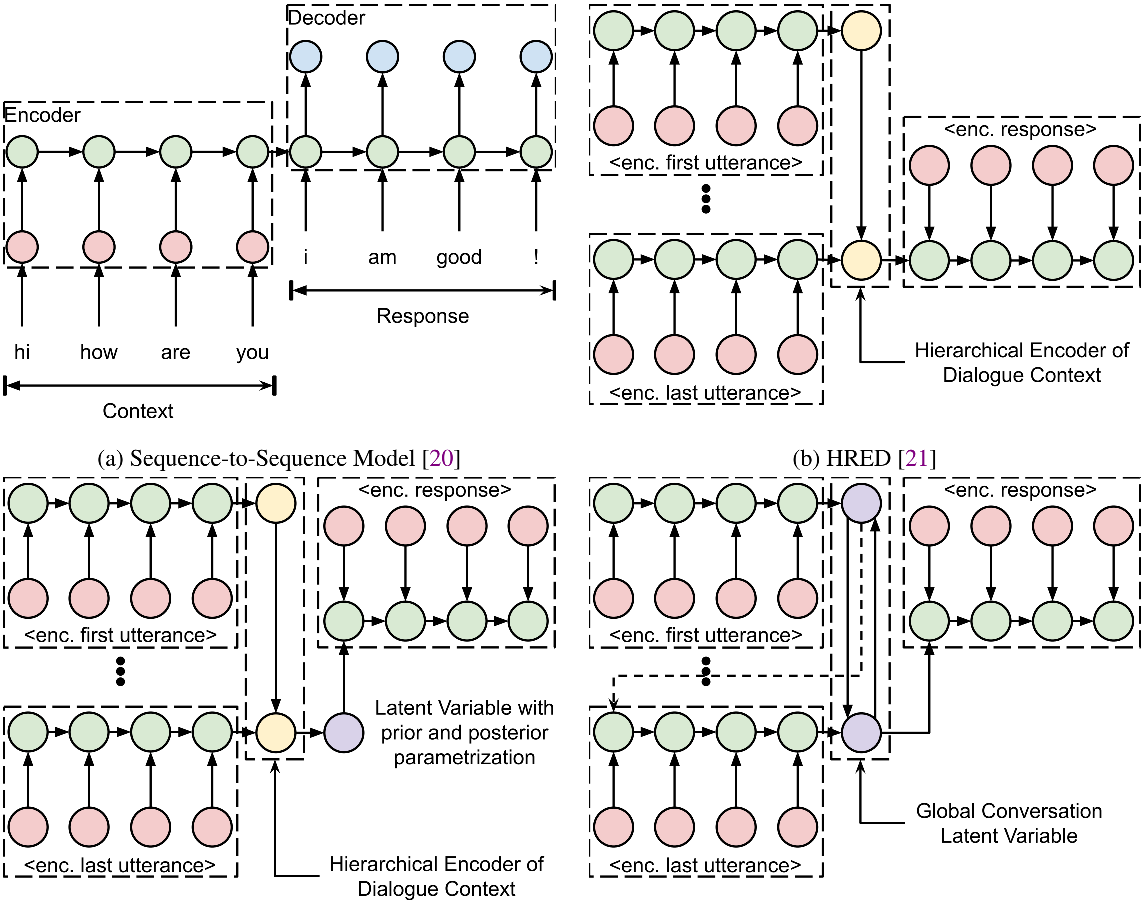
4.2.1 Dialogue Modelling

To investigate the usability of anomaly detection for dialogue evaluation, we consider four neural network models for dialogue modeling. These approaches tackle conversations by first encoding the input context and using that representation by decoding it into the response. *While this is not the same as autoencoders, we can use the loss measuring the correctness of mapping the context to the reply in the same manner as a reconstruction loss.* In this subsection, we concisely present the models used for this study. For more detail on each of the approaches, we would forward the reader to the appropriate reference, during each of their presentations.

The first model (shown in Figure 4.1(a)) we consider is a recurrent sequence-to-sequence approach, as described by Vinyals and Le [20]. It models a dialogue as a sequence of pairs of query and response, i.e., it considers a response as related only to the last utterance before it. The context is encoded using a recurrent neural network (RNN), and another RNN decodes the representation into the response. Cross-entropy acts as a reconstruction loss measuring how well the utterance maps to the context.

Next is Hierarchical Recurrent Encoder-Decoder (HRED) by Serban et al. [21], which builds upon the sequence-to-sequence (Seq2Seq) approach by considering multiple utterances from the context

[1] N is a hyperparameter



(a) Sequence-to-Sequence Model [20]

(b) HRED [21]

(c) VHRED [22]

(d) VHCR [23]

Figure 4.1: Architecture diagrams of the four approaches used in this section.

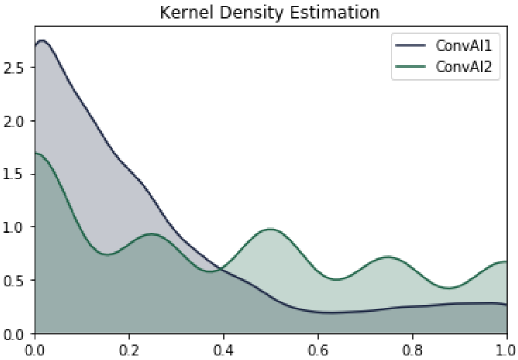
(displayed in Figure 4.1(b)). It does so by using a third RNN. The context utterances are each encoded using an RNN, and then encoded together one vector representation by the additional RNN. The rest is as in the sequence-to-sequence approach described earlier.

Thirdly, Serban et al. [22] propose an extended version of HRED, a Hierarchical Latent Variable Encoder-Decoder (VHRED), by adding a latent variable at the decoder that parametrizes the context (visualized in Figure 4.1(c)). Kullback-Leibler (KL) divergence provides measures of the reconstruction between the original context representation and its latent variable version. This way, the approach can model hierarchically-structured sequences in a two-step generation process-first sampling the latent variable, and then generating the output sequence-while maintaining long-term context.

Finally, Park et al. [23] report that VHRED sufferers from a degeneration of the latent variable, which renders the model to behave almost like an HRED. They introduce a global conversation latent variable such that it is responsible for generating each of the utterances of the dialogue rather than capturing the whole context post-factum. The method's architecture is illustrated in Figure 4.1(d).

To train all the models, we use the Cornell Movie-Dialogs Corpus [83]. It has 220,579 conversations and a total of 304,713 utterances. The training is done by iterating over each dialogue turn and considering the full query context. The first sequence-to-sequence approach is using only the last dialogue turn as a context.

4.2.2 Dialogue Datasets



(a) Distribution of annotator scores.

Feature	ConvAI1	ConvAI2
# Dialogues	2154	2237
Avg # Utterances	13.9	18.1
Avg # Words per Utterance	7.3	8.2
Task	Topic discussion	Person impersonation

(b) Key features of the dialogue datasets.

Figure 4.2: Overview of the ConvAI1 and ConvAI2 datasets. We see that the majority of dialogues are evaluated as low quality. Only dialogues with three or more utterances were considered as part of this work.

We use the data gathered during the ConvAI1[2] [1, 84] and ConvAI2[3] [79, 85] challenges. The organizers invited competitors to develop dialogue systems that had to address specific tasks. For ConvAI1, the participating systems needed to be able to converse about a topic. In the other competition, the chatbots had to engage in a small-talk while impersonating a pre-defined personality profile ("persona"). In both cases, human annotators evaluated the capability of the dialogue systems to converse by interacting with them and giving a score at the end. For both competitions, the scoring is on dialogue level. In Figure 4.2, we present some additional details about the data. However, we do not evaluate the two challenges specifically (topic discussion and role acting). Instead, we aim at general open-domain dialogue evaluation, which implies relevance, coherence, and fluency of the utterances.

4.2.3 Scoring

As presented in 2.4.1, the cross-entropy loss function will act as a reconstruction loss to detect anomalies. For obtaining the scores, the dialogues presented in subsection 4.2.2 go through the same iterative manner described in subsection 4.2.1. After that, the scores are averaged on the dialogue level to obtain a single value that summarizes the whole conversation.

Cross-entropy is defined as:

[2] http://convai.io/2017/data/
[3] http://convai.io/data/

Dataset	**ConvAI1**		**ConvAI2**	
	r	ρ	r	ρ
Seq2Seq	0.2150	0.3006	0.3444	0.4892
HRED	0.1869	0.2832	0.3469	0.4876
VHRED	0.2210	0.3009	0.3384	0.4885
VHCR	0.2249	0.3037	0.3408	0.4888

Table 4.1: Pearson's correlation coefficients, r, and Speaman's correlation coefficients, ρ, on the two dialogue datasets' human scores and cross-entropy scores. All of the scores are with a confidence of $p < 0.0001$

$$\begin{aligned} L &= \frac{1}{T}\sum_{t=1}^{T} l_t \\ l_t &= -w_r(\sum_{v=1}^{V} y'_v \log(y_v)) \end{aligned} \tag{4.1}$$

where t stands for the t-th token in the response, y'_v, and y_v are the true and the predicted words from the vocabulary (V), respectively, w_r are weights used for ignoring padding tokens in a sequence. All of the scores obtained from a single model applied to a dataset undergo a rescaling such that the maximum will have a value of 1.0.

4.3 Evaluation

In this section, we will analyze the dialogue datasets, ConvAI1, and ConvAI2, separately for possible correlations between the cross-entropy values exhibited from each of the models and the respective annotator score. The results are summarized in Table 4.1.

The first immediate observation is that all of the models across the two datasets demonstrate a significant positive correlation with the scores from the human annotators. The result is contrary to the initial expectation for the following reason. Cross-entropy measures the models' ability to reconstruct a response from the given query context. Thus, the higher the loss function's value is, the more difficult it is for the model to relate the input to the response. The positive correlation states that as the annotator's score increases, so does the cross-entropy. Ideally, the correlation between the two variables should be negative, since the models used training data with proper examples and, thus have difficulties to process anomalous conversations from dialogue systems. Then, the outlier exchanges will be lowly evaluated by the human annotators, and the models should have a comparatively higher loss score.

Furthermore, all of the approaches appear to have a shared understanding and perspective of the conversations because they are demonstrating a very similar correlation with the annotators' scoring. The sequence-to-sequence approach is also on par with the others, which is noteworthy because unlike the others, it cannot capture long-term dependencies in dialogues. Thus, long-term context appears to be not necessary for the scoring of these dialogues by the annotators.

We see that in Figure 4.2 that the dialogue scores by the annotators have a non-uniform distribution. Thus, we set to investigate if there are any patterns within the various quality subgroups. For that purpose, we split the dialogues into five equal-width bins based on the minimum (0.0) and maximum (1.0) values for the human annotator scores. All of the sub-groups that exhibit somewhat negative correlation coefficients are in Table 4.2.

For the dialogues in ConvAI1, we discover that all of the models exhibit a very weak negative correlation in the quality scores between 0.4 and 0.8. The considerably lower amount of examples in the groups with higher quality contributes to low confidence estimates. Nevertheless, this discovery hints that there is limited potential in using anomaly detection for dialogue quality evaluation.

Meanwhile, for the conversations in ConvAI2, we identify stronger than in ConvAI1 negative correlations with the top-most in terms of quality samples. The dialogues in the quality range between 0.8 and 1.0 have negative Pearson's and Spearman's correlation coefficients. These samples provide further evidence to the potential of having an anomaly detection perspective on the issue.

Model	**Dataset**	**Quality Range**	r ($p \leq$)	ρ ($p \leq$)
Seq2Seq	ConvAI1	$[0.4, 0.6)$	0.0141 (0.8087)	-0.0513 (0.3791)
Seq2Seq	ConvAI1	$[0.6, 0.8)$	-0.0093 (0.9309)	0.0941 (0.3776)
Seq2Seq	ConvAI2	$[0.8, 1.0]$	-0.0093 (0.9309)	-0.0093 (0.3791)
HRED	ConvAI1	$[0.4, 0.6)$	0.0145 (0.8039)	-0.0514 (0.3783)
HRED	ConvAI2	$[0.8, 1.0]$	-0.2493 (0.0001)	-0.2778 (0.0001)
VHRED	ConvAI1	$[0.4, 0.6)$	0.0093 (0.8737)	-0.0546 (0.349)
VHRED	ConvAI1	$[0.6, 0.8)$	-0.0097 (0.9279)	0.0984 (0.3562)
VHRED	ConvAI2	$[0.8, 1.0]$	-0.2613 (0.0001)	-0.282 (0.0001)
VHCR	ConvAI1	$[0.4, 0.6)$	0.0106 (0.8559)	-0.0507 (0.3843)
VHCR	ConvAI1	$[0.6, 0.8)$	-0.0196 (0.8546)	0.0958 (0.3689)
VHCR	ConvAI2	$[0.8, 1.0]$	-0.2609 (0.0001)	-0.2841 (0.0001)

Table 4.2: Selected sub-groups with negative correlation coefficients. The omitted groups have positive correlations aligned with the results from Table 4.1.

4.4 Summary

On a high level, we saw that the method is unfit for replacing human annotators. However, when we consider only various quality sub-groups of the data, the models demonstrate an expected negative correlation and show some promise for using their loss function outputs for detecting anomalous conversations.

Overall, the limited ability to generalize or, otherwise, the insignificant amount of training data are obstacles for using outlier detection methods for evaluating dialogues. As future work, we would focus in this direction, so that models can better generalize and be able to demonstrate consistent behavior across various domains, thus, successfully assessing dialogue quality.

CHAPTER 5

Language Models as Evaluators

> Science is organized knowledge. Wisdom is organized life.
>
> *Immanuel Kant*

In Chapter 4, we discussed an anomaly-detection driven approach, that unfortunately suffered from limited generalization abilities as we saw. Hence we turn our attention to pre-trained language models, that have been exposed to copious amounts of text data, and it is difficult to "surprise" them with something they have not seen.

Going back to the analogy of how our indispensable human annotators evaluate a dialogue, they do not use an explicit reference or necessarily seek word overlap between context and response (or the lack of it). Rather, their assessment bases itself on experience with the language and the implicit knowledge they have about it. The core principle of statistical language models (LM) is to capture and reproduce these properties. LMs have proven themselves invaluable in state-of-the-art approaches in natural language processing, and natural language understanding [2–4, 27].

Research Question 2

Can language models indicate the quality of a conversation?

Thus, the main aim of this chapter[1] is to investigate their usability as means for evaluating dialogues since they do not need a reference or supervision. We demonstrate that there is a significant positive correlation between the predictions of language models and human evaluation scores. Furthermore, we provide insights into the inner-workings and behavior of language models in the dialogue context.

The current chapter of the thesis is using the following publication as a basis:

> **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2020. Language Model Transformers as Evaluators for Open-domain Dialogues. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6797–6808, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[1] Code and resources to reproduce the results are available on the following link: https://github.com/SmartDataAnalytics/transformers_dialogue_evaluators

We investigate three different approach for language models for the purpose of evaluating dialogues. Each of them is considered pivotal in the research of LMs. However, unlike the anomaly detection approach from the last chapter, they require no adaptions.

5.1 Background

In this section, we shortly revisit related work that focuses on dialogue evaluation. Furthermore, we revisit language-model-based transformers, as an extension to the discussion in Section 2.3 and recent advances in this particular set of approaches.

Earlier works ([51, 52]) that we mentioned in 3.1 are can be seen as direct competitors of the current approach. However, all of them require a ground truth to provide evaluation output for a dialogue.

On another note, Kann et al. [86] suggest a sentence level fluency metric derived from the perplexity score of a language model given a sentence without involving any references. Their results demonstrate significant positive correlations with human annotators.

In Section 3.1, we discussed two works by Mehri & Eskenazi [57] and Sai et al. [55] that can be considered related to what we propose here. In the case of the former, the authors use a fine-tuned RoBERTa model on dialogue data together and its masked language modelling objective. In addition, they employ a dialogue retrieval approach to rate the suitability of the responses. Regarding the work of the latter, the research team uses BERT's next sentence prediction that is fine tuned to predict the next utterance instead on dialogue data. Unlike these two works, our approaches are applied as they are without any adaptions or modifications. It shall be noted, that all three works, (including ours), have had their articles published in the same year.

5.2 Methodology

In this section, we report on the used datasets for assessing the usability of transformer language models for evaluating dialogue quality, introduce the used approaches in greater detail and describe their relevance to the task at hand.

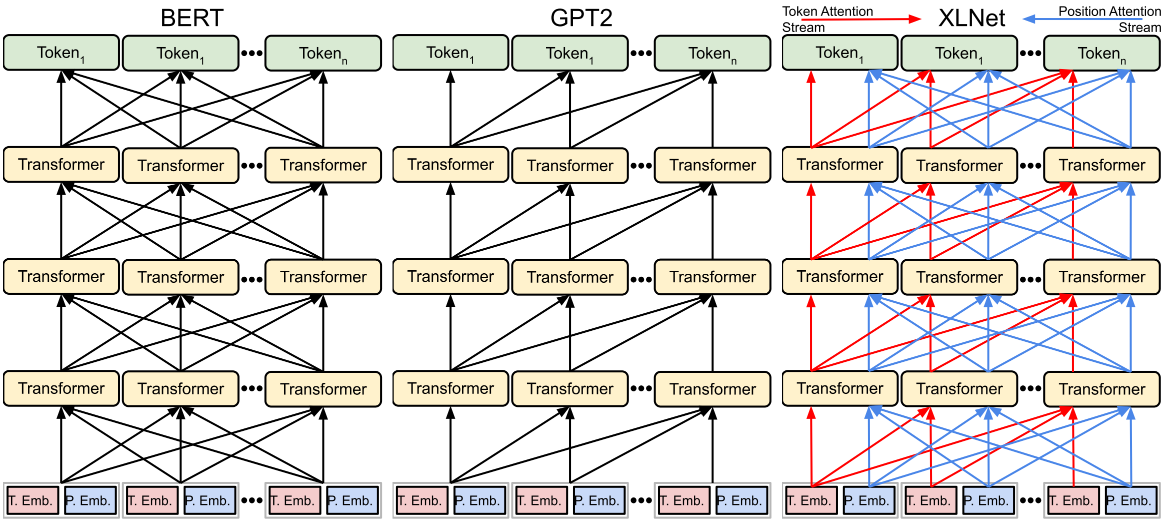
5.2.1 Datasets

We use the same datasets as in the previous Chapter 4 - ConvAI1 [1, 84] and ConvAI2 [79, 85] challenges. We kindly forward the reader to Section 4.2.2 for mode details and information on the two datasets.

5.2.2 Language Model Evaluators

In Section 2.3, we presented an introduction into transformer-based language models. In the current subsection, we will provide more details about three of those architectures, and how we use them for conducting this study. Our main goal is to use the LM to assign a probability to the utterances in a conversation. We used HuggingFace's Transformers[2] [87] for implementation and pre-trained weights of transformer-based language models.

[2] https://github.com/huggingface/transformers



(a) BERT [2]

(b) GPT2 [3]

(c) XLNet [4]

Figure 5.1: ARchitecture of the RUBER approach. On the left, we have the high-level

Since intuition dictates that responses are dependent on their preceding context, we condition the target reply on its history to measure its relevance. Kann et al. [86] showed how language models could serve as good sentence-level fluency indicators. Thus, the calculated probability from the transformer-based LM can serve as a combined score for fluency and coherency. The following LMs are used in this chapter:

1. As previously mentioned, BERT [2] is using two language modeling objectives: masked language modeling (MLM) and next sentence prediction (NSP). MLM provides no viable way for computing the probability of a target response because it originally substitutes only a random subset of tokens. Thus, there is no consistent and deterministic way to use masked language modeling for assigning a probability score to a response given its context. However, BERT's next sentence prediction is an excellent approach for the current task. It can judge if an utterance is the next one given its contextual predecessor. Thus, we pair up the sequentially appearing sequences in a conversation and compute a probability score for the second reply:

$$P(u_2|u_1) = P(t_{21}, t_{22}, ..., t_{2n}|t_{11}, t_{12}, ..., t_{1m}) \tag{5.1}$$

   ,

   where $P(u_2|u_1)$ is the probability score of the target response, while $(t_{11}, t_{12}, ..., t_{1m})$ and $(t_{21}, t_{22}, ..., t_{2n})$ are the tokens belonging to the query and response utterances prospectively.

2. The approach of GPT2 [3] is the standard language model approach that factorizes the joint probability over the sequence tokens $(t_1, t_2, ..., t_n)$ as a product of the conditional probabilities [27]:

$$P(x) = \prod_{i=1}^{i} P(t_n|t_1, t_2, ..., t_{i-1}) \tag{5.2}$$

In our problem domain, we need to consider two consecutive sequences and capture the coherence between them. Thus, we concatenate them into one, where the context appears first and is then followed by the second utterance. We then compute the joint probability for the second part conditioned on the past:

$$P(x) = \prod_{i=m+1}^{m+n} P(t_{m+n}|t_i, t_{n+1}, ..., t_{m+n-1}) \tag{5.3}$$

where m is the length of the context, and n is the length of the target utterance.

3. XLNet [4] follows the same general language model approach as GPT2, however, with some additions to its training objective and neural network architecture. First of all, unlike GPT2, XLNet optimizes the model over a sequence w.r.t. all possible permutations of the factorization orders rather than each one separately. Secondly, compared to conventional neural transformers, XLNet adds one more attention stream that includes the positional information of the target token but excluding the content to maintain the autoregressive properties. To compute probabilities for the utterances, we follow the same procedure as described above for GPT2.

In this work, we use a set of hyper-parameter configurations for each of the three language models. We present them in Table 5.1.

Name	**Details**
bert-base-uncased	12-layer, 768-hidden, 12-heads BooksCorpus English Wikipedia
bert-large-uncased	24-layer, 1024-hidden, 16-heads BooksCorpus & English Wikipedia
gpt2	12-layer, 768-hidden, 12-heads news, Wikipedia, fiction books
gpt2-medium	24-layer, 1024-hidden, 16-heads news, Wikipedia, fiction books
gpt2-large	36-layer, 1280-hidden, 20-heads news, Wikipedia, fiction books
xlnet-base-cased	12-layer, 768-hidden, 12-heads same as BERT + news
xlnet-large-cased	24-layer, 1024-hidden, 16-heads same as BERT + news

Table 5.1: Hyper-parameter configurations (number of layers, size of the hidden state, number of attention heads) of the models and used corpora to pre-train them. Source: `https://huggingface.co/transformers/pretrained_models.html`

5.2.3 Scoring

In Equations 5.2 and 5.3, we showed how language models compute a probability score for a whole sequence. However, as an aggregated score over the tokens, it is losing the initial probabilistic distribution over the tokens. Furthermore, since we are dealing with dialogues, i.e., a sequence of utterances, we need to perform two levels of aggregation. The first level is an aggregation of the word tokens within an utterance, while the second is the done while aggregating over the utterances.

Thus, we investigate other possible ways to derive an aggregated score over the word tokens and over the utterances within a dialogue. Besides a product of probabilities, we also look into a sum and an unweighted average, which capture the length of the sequences (utterance or dialogue), which might prove beneficial for a correlation study with human annotators. We normalize all of the scores such that they range between 0.0 (population minimum) and 1.0 (population maximum).

For GPT2 and XLNet, our experiments show that the following formulation correlates the highest with human annotator scores:

$$lm\_dialog\_score = \sum_{u=1}^{Utterances} \left(\frac{\sum_{w=1}^{Words} P_{(w=w)}}{\#Words} \right) \tag{5.4}$$

In addition, we experimented with the following formulations of a LM-based evaluation score:

$$lm\_dialog\_score = \sum_{u=1}^{Utterances} \left(\sum_{w=1}^{Words} P_{(w=w)} \right) \tag{5.5}$$

$$lm\_dialog\_score = \frac{1}{\#Utterances} \sum_{u=1}^{Utterances} \left(\frac{\sum_{w=1}^{Words} P_{(w=w)}}{\#Words} \right) \tag{5.6}$$

$$lm\_dialog\_score = \frac{1}{\#Utterances} \sum_{u=1}^{Utterances} \left(\sum_{w=1}^{Words} P_{(w=w)} \right) \tag{5.7}$$

$$lm\_dialog\_score = \prod_{u=1}^{Utterances} \left(\sum_{w=1}^{Words} P_{(w=w)} \right) \tag{5.8}$$

$$lm\_dialog\_score = \prod_{u=1}^{Utterances} \left(\frac{\sum_{w=1}^{Words} P_{(w=w)}}{\#Words} \right) \tag{5.9}$$

$$lm\_dialog\_score = \prod_{u=1}^{Utterances} \left(\prod_{w=1}^{Words} P_{(w=w)} \right) \tag{5.10}$$

The correlation coefficients between these aggregations scores and the human annotation are either of low values, are insignificant (low p-value), or both.

5.2.4 Baseline

We take RUBER from Tao et al [52] as a baseline. The approach initially employs two components that perform two functions. The first one is to calculate a resemblance score using word vector pooling and references. We aim for an unreferenced evaluation approach akin to a human evaluator. Thus, we use only the second component of the method. This second component can calculate a relevance score for a given response based on its preceding context. It uses a bidirectional GRU network and negative sampling. To reproduce as best as possible the original results of RUBER, we sample 1,449,218 pairs of sequential utterances from the OpenSubtitles dataset [88].

5.3 Evaluation

In this part of the chapter, we will conduct a correlation analysis between the calculated probabilities from the LM and the scores given to dialogues by human evaluators. We provide a closer look at

Dataset	**ConvAI1**		**ConvAI2**	
	r	ρ	r	ρ
bert-nsp-d-sum	0.169	0.273	0.205	0.490
bert-large-nsp-d-sum	0.172	0.277	0.205	0.485
gpt2-u-avg-d-sum	-0.027	0.068	0.152	0.323
gpt2-md-u-avg-d-sum	-0.005	0.069	0.144	0.325
gpt2-lg-u-avg-d-sum	-0.038	0.048	0.127	0.325
xlnet-u-avg-d-sum	0.068	0.157	0.206	0.435
xlnet-lg-u-avg-d-sum	0.087	0.169	0.225	0.437
RUBER-U	0.154	0.129	0.013	-0.005

Table 5.2: Pearson's r, and Speaman's ρ, correlation coefficients on the two dialogue datasets' human scores and various aggreggated scores from the language models. "u-avg-d-sum" stands for averaged probabilities on utterance level and then summed up on conversation level. Most of the scores are with a confidence of $p <= 0.001$. Exceptions are GPT2-medium and GPT2 in ConvAI1 with 0.812 and 0.212 respectively, as well as, RUBER-U for ConvAI2, both r and ρ, with 0.5309 and 0.8166, respectively.

some auxiliary model outputs as well.

5.3.1 Quantitative Assessment

In Table 5.2, we report the noteworthy Pearson's and Spearman's correlation coefficients between the aggregated probability scores and the evaluations of the dialogues.

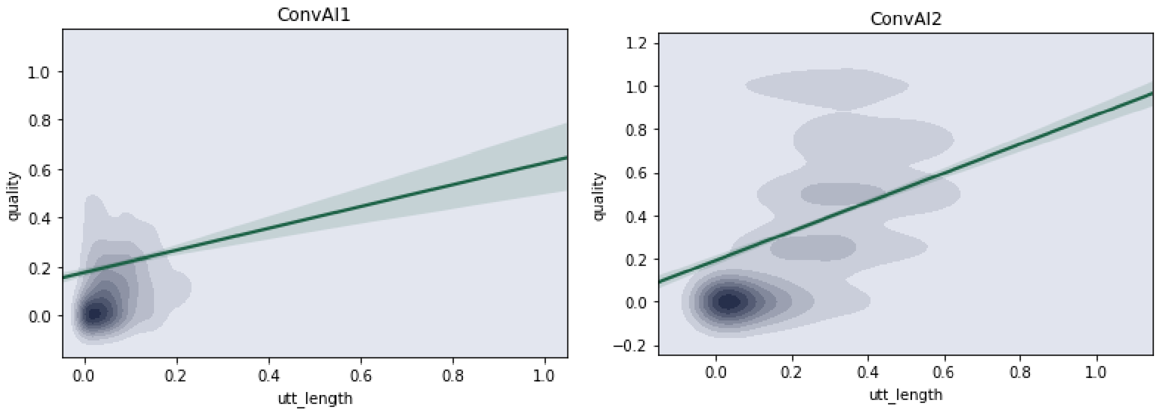
The immediate observation of using language models as dialogue evaluators shows that there are gaps in terms of performance between the three different approaches. Most evident is the difference between BERT and the others. Its next sentence prediction objective explains this behavior. Unlike the other two, BERT takes the most structured approach to modeling two sequences. It recognizes the two utterances as separate and captures their information as a whole. Thus, when we compare it to GPT2 and XLNet, it has the advantage of not needing score aggregation on utterance level, because it produces a probability for the whole sentence rather than word for word.

Also, there is a smaller difference in performance between GPT and XLNet. First of all, they share a core foundation as autoregressive language models, thus are more similar to each other than BERT, which also explains their overall behavioral similarity. However, XLNet has a structural improvement in its architecture. Unlike GPT2, it also encodes the positional information of the target token. Thus, similarly to BERT, it can capture more information about a sequence and consequently have a better correlation score.

Additionally, we investigate the effect of model size. The difference in correlation coefficients between the hyperparameter configurations is marginal and, in one of the cases, even non-existent. The most evident example is the spectrum displayed by the three GPT2 settings. Ultimately, we can conclude that smaller models perform similarly at a much smaller energy cost.

In regards to score aggregation, all the approaches unanimously show that averaging on utterance level and summing up the whole conversation is the most informative for dialogue evaluation. At the same time, the using a product or an unweighted average produce correlation coefficients very close to zero and with an extremely low significance (e.g., $p - value$ ranging from 0.4 to 0.8). The behavior indicates that while utterance length is insignificant, the duration of the conversation strongly dictates its quality score.

5.3.2 Qualitative Assessment



(a) ConvAI1

(b) ConvAI2

Figure 5.2: Regression plots showing the relation between quality score and utterance length in the ConvAI1 and ConvAI2 datasets. The shaded area around the line represents a confidence interval.

In Figures 5.2(a) and 5.2(b), the regression models show the interaction between the annotator quality score and the length of a conversation in ConvAI1 and ConvAI2, respectively. In both cases, the regression shows a positive trend that the longer a dialog is, the better its assessment is. We also see that in the case of ConvAI1, the confidence area is much wider than in ConvAI2. This behavior further supports the results in Table 5.2, where the language models have considerably lower correlation coefficients for ConvAI1.

Furthermore, we manually investigated short conversations from both datasets that also have low quality. Many of the short dialogues show that the system would indeed perform poorly by not responding at all, or the first couple of utterances would be not diverse or even the same. Thus, the annotator would terminate the session and evaluate the dialogue with a low score. In contrast, conversations that were more interactive and had longer duration also performed better in their assessment.

5.3.3 What Would a Language Model Say?

In this subsection, we report the correlation scores between the maximum probabilities for each token and the annotator scores. The intuition is that besides being renown for advancing the state-of-the-art in various NLP benchmarks, language models are prominent for being capable generators of natural language. Furthermore, Hendrycks and Gimpel [40] have demonstrated that the maximum class probability of a neural network classifier tends to output lower values for samples that are out of distribution. Thus, we set to investigate whether the predicted maximum classes of language models can also indicate the quality of dialogues.

Although there are some studies [89] demonstrating BERT generating text, we will not consider it in this part of the work due to the nature of its masked language modeling, which does not aim at generating text. Considering GPT2 and XLNet, we look into what are the most likely words they predict for each token of the sequence instead of the original ones.

For the context of dialogue evaluation, it means that on average *max* scores should be higher for fluent and coherent text like the one used for pretraining the language models. At the same time, erroneous samples should have lower maximum probabilities.

Firstly, we investigate the quantitative relation of the *max* scores to human annotator scores. Similarly to what we did in Section 5.3.1, we have calculated the aggregated probability scores for the most likely words according to the language models (shown in Equation 5.11).

$$lm\_dialog\_score_{max} = \sum_{u=1}^{Utterances} \left(\frac{\sum_{w=1}^{Words} P_{(w=w_{max})}}{\#Words} \right) \tag{5.11}$$

Original Context	**Original Response** (as in dataset)	**Generated Response** (generated by transformer)
"Wow! Are you man or woman?"	"I am! i am a woman."	" ’m a I am a man! I"
"How nice! Do you have a boyfriend?"	"I do not. i am a single mom."	" ’m . . I am a virgin woman. i"
"What do you mean?"	"granted the right to accept only one religion"	"anted, fact to be or the of"
"Do you know Utrecht?"	"granted the right to accept only one religion"	"ind, title to use donations Dutch application"

Table 5.3: Sample dialogue exchanges as originally seen in the ConvAI1 and ConvAI2 datasets together with alternative responses generated by GPT2 by just taking the most likely word. Coherent examples induces the language models to generate also good response. The top two examples have high human annotator scores, while the bottom two are rated lowly.

Dataset	**ConvAI1**		**ConvAI2**	
	r	ρ	r	ρ
gpt2-u-avg-d-sum	0.133	0.261	0.193	0.477
gpt2-md-u-avg-d-sum	0.144	0.263	0.196	0.476
gpt2-lg-u-avg-d-sum	0.146	0.267	0.196	0.477
xlnet-u-avg-d-sum	0.157	0.263	0.211	0.471
xlnet-lg-u-avg-d-sum	0.137	0.251	0.209	0.475

Table 5.4: Pearson’s correlation coefficients, r, and Speaman’s correlation coefficients, ρ, on the two dialogue datasets’ human scores and various aggregated scores for the *max* word instead of the target. "u-avg-d-sum" stands for averaged probabilities on utterance level and then summed up on conversation level. All of the scores are with a confidence of $p < 0.001$.

We present the results in Table 5.4. When compared to the analogous results in Table 5.2, we see that GPT2 and XLNet demonstrate noticeably higher correlation coefficients, especially for the dialogues in the ConvAI1 dataset. This discrepancy suggests that for some of the cases, the models can generate text that would fit better into the conversation. Since, ConvAI1 and ConvAI2 happened before the introduction of transformer-based language models, it is save to assume that the participating systems are inferior.

In Table 5.3, we present some short sample conversations together with a generated text by a language model. The top two examples have high scores by the human annotators, while the rest are of low quality. The model can reconstruct sensible responses that make sense and are still different from the original reply. On the other hand, whenever there is an incoherent conversation like the third and fourth examples, GPT2 and XLNet are not able to recreate a response that is either somewhat fluent or related to the current context. Another peculiarity is that the language model possesses in a sense, common knowledge. This is demonstrated by the fourth example, while in the preceding utterance, we see Utrecht, a Dutch city, and the model is then induced to predict "Dutch" as one of the response tokens.

5.4 Summary

In this study, we investigated whether transformer-based language models can evaluate dialogues in terms of coherency and fluency. Overall, Pearson's and Spearman's correlation coefficients demonstrate that BERT, GPT2, and XLNet can indicate a conversation's quality without any additional supervision or reference. While, in their core, the three use the same approach, transformers, they have further structural modifications that set them apart when considered for the current problem domain.

GPT2 performs worst due to its standard language modeling approach that incorporates the least structural information about a sequence. XLNet achieves an improvement in terms of its correlation score by taking advantage of additional positional information when predicting a target token. Finally, BERT's next sentence prediction approach delivered the highest performance thanks to its structured approach in regards to separate utterances.

While LM-based dialogue evaluators cannot yet replace human annotators, they have additional value when compared to word-over metrics like BLEU or ones that use word-embeddings. Although they cannot completely replace human evaluators, They can support as weak indicators for quality. Additionally, we have shown that they can perform better than competing approaches like the unreferenced component of RUBER.

Furthermore, the autoregressive language models, GPT2 and XLNet, demonstrate an excellent initial aptitude for conducting dialogues. They can provide alternative responses that are also coherent with the context of a discussion.

CHAPTER 6

Proxy Indicators for Dialogue Quality

> Never discourage anyone...who continually makes progress, no matter how slow.
>
> *Plato*

In the last Chapter, Chapter 5, we demonstrated how transformer-based language models with different training objectives can effectively evaluate dialogue systems. However, they all have one major disadvantage - they offer only a single score that evaluates only the overall quality, but provides no targeted insight about the specific dialogue features like fluency or coherency. Hence, we are need to find a way to measure those features independently.

Once again, we revisit the analogy of how our invaluable human annotators evaluate a dialogue, instead of using a sample response, they usually would detect issues with conversations based on natural criteria like fluency or coherency. These are related to a specific skills that people acquire while learning any language. For decades, NLP researchers have tried to teach such skills to computers. Usually, this is achieved by the introduction of specific benchmarks.

Research Question 3

Are standard NLP tasks helpful with the evaluation of dialogues?

Thus, the main goal of this chapter[1] is to investigate the usability of standard benchmarks as means for evaluating dialogues since they do not need a reference or supervision. We demonstrate that there is a significant positive correlation between the predictions on benchmarks and human evaluation scores. Furthermore, we demonstrate a composable metric based on those predictions that allows focus on various criteria.

The current chapter of the thesis is using the following publication as a basis:

Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2021. Proxy Indicators for the Quality of Open-domain Dialogues. In Proceedings of the 2021 Conference on Empirical

[1] Code and resources to reproduce the results are available on the following link: https://github.com/SmartDataAnalytics/proxy_indicators

Methods in Natural Language Processing, pages 7834–7855, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

We investigate whether natural language processing (NLP) tasks can serve as proxy indicators for a conversation's quality. For that purpose, we use a fine-tuned BERT [2] model trained on the GLUE benchmark [5]. GLUE provides a comprehensive evaluation of general language understanding. We demonstrate that a few of the tasks exhibit a limited potential of serving as proxy indicators. The rest shows negative results.

6.1 Background

More recently, Ghandeharioun et al. [69] propose a framework that uses self-play and two NLP tasks as an additional source of knowledge to evaluate dialogues in a multi-turn mode scenario. They perform an ablation study using sentiment and natural language inference as proxy supervision to see whether their system can better approximate human judgment. Their work shows that dialogue systems can benefit from using them. Also, Welleck et al. [90] frame the dialogue consistency issue as a natural language inference problem and propose the DialogueNLI dataset. Its purpose is to benchmark a model's ability to select relevant utterances relative to a given context.

6.2 General Language Understanding Evaluation

In Chapter 3.4, we have presented the GLUE benchmark in detail and its respective tasks. It serves as the foundation of NLP tasks that will act as indicators for the various dialogue criteria. Hence, we kindly forward the reader to read it for more details.

6.3 Methodology

6.3.1 Dialogue Datasets

To evaluate the ability of a deep-learning model trained on GLUE to indicate the quality of dialogues, we use the English datasets (TopicalChat, PersonaChat) provided by Mehri & Eskenazi [57]. They train a few different dialogue system models and use different sequence generation techniques to generate responses for certain dialogue contexts. The researchers then evaluate 660, in total, dialogue contexts and responses according to six criteria:

- **Understandable** *(0 - 1)* - Can a user understand the final response given the context?
- **Natural** *(1 - 3)* - Does the user find the response like something that a real person would say?
- **Maintains Context** *(1 - 3)* - Is the response part of the established flow in the conversation?
- **Interesting** *(1 - 3)* - Is the utterance boring or does it contribute to the discussion?
- **Uses Knowledge** *(0 - 1)* - Given a background knowledge base, how well does the response make use of it or relate to it?

- **Overall Quality** *(1 - 5)* - Considering the previous criteria, what is the general perception of the quality of the utterance?

Each of the conversations context has been evaluated by three annotators. The authors of the dataset report that most inter-annotator agreement or correlation scores for the criteria are above 0.4, which suggest a moderate to strong agreement.

For further details about the dataset, we forward the reader to the original work of Mehri & Eskenazi [57].

6.3.2 BERT as a Proxy Indicator for Dialogue Quality

Since the GLUE benchmark is about *general language understanding*, we are interested to know whether a model trained on it can indicate the quality of the dialogue. To conduct the investigation, we use BERT [2] and its fine-tuned models on the GLUE benchmark [91, 92]. We use the version with 110M parameters. For each investigated GLUE task, there is a separate copy of the whole model trained to solve that specific problem. While we did not train the models ourselves, the inference is less demanding. It takes about 30 minutes on a laptop with an eight-generation Intel i7 CPU.

For encoding the text sequence, we use several instances of BERT (shown in Figure 6.1), a

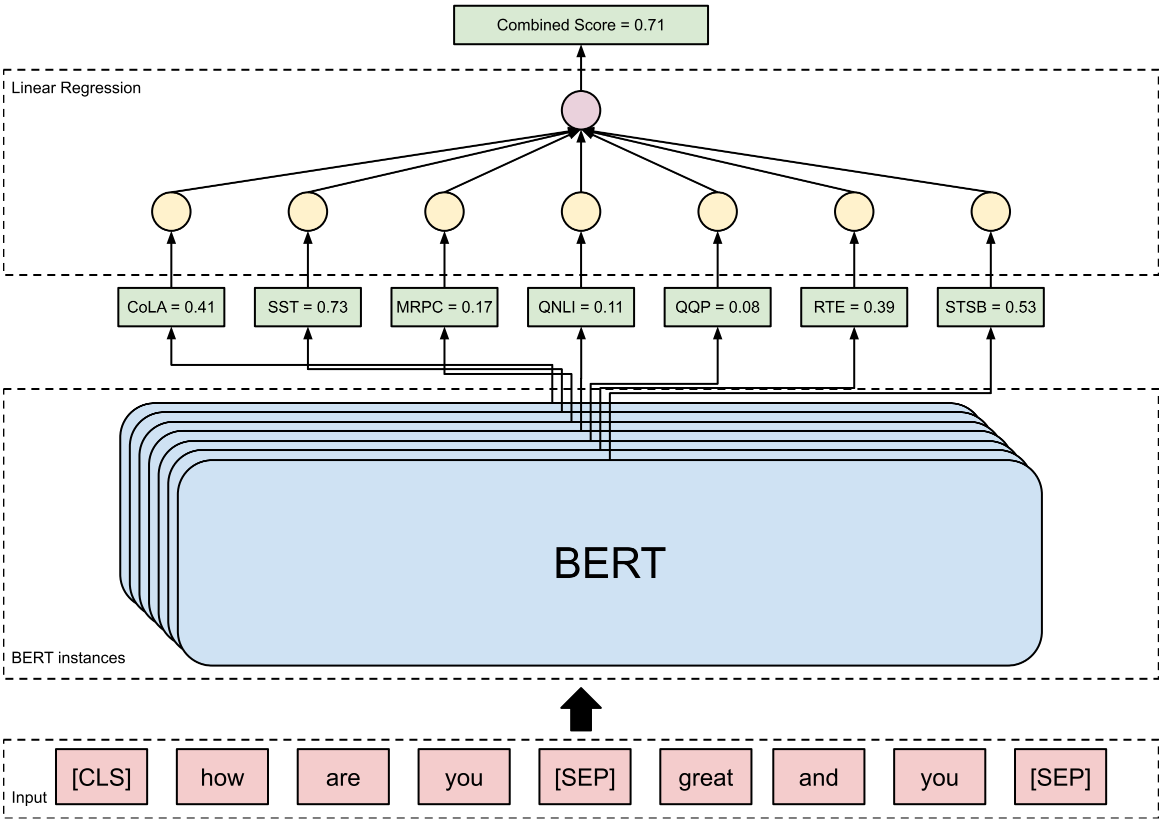


Figure 6.1: An example demonstrating the usage of multiple BERT instances for the various GLUE tasks. In additon, we see all the scores combined into one using a linear regression.

pre-trained bidirectional transformer encoder language model. The pre-training has been done using two unsupervised tasks: masked language modeling and next sentence prediction. This way, it can learn a contextualized semantic representation of the input text usable for downstream tasks. BERT can create a vector encoding for a whole sequence by always inserting a control token, $[CLS]$, at the beginning. For the case of pair-wise sentence tasks, e.g., next sentence prediction, it uses an additional control token, $[SEP]$, between the two sentences to distinguish them.

When fine-tuned for a specific task, the pre-trained language model weights are reused. In addition, a layer is added to act as a transformation from BERT's semantic representation to the space of the target variable, e.g., the classes of RTE or CoLA.

6.3.3 Scoring

For obtaining model predictions, the dialogue data is provided as input in three possible ways: 1. single utterance, 2. a dialogue context and a response, or 3. related facts to a conversation and a response. Depending on the GLUE task, the model can give **four different types of output scores**:

Single-sentence classification output provides softmax output for CoLA and SST-2. Given the contextualized semantic representation of a single utterance from the dialogue U the probability whether it is linguistically acceptable or with a positive sentiment is:

$$\begin{gathered} P_r(c_{task}|U) = \text{softmax}(W_{task}^T \cdot U), \\ task \in \{\text{CoLA}, \text{SST-2}\} \end{gathered} \tag{6.1}$$

where W are the task-specific weights, c is the output class for the target task.

Pairwise text similarity outputs a similarity score, for the STS-B task, between a pair of a context or fact and a target response from the same dialogue C (or F for a fact) and R, concatenated and jointly encoded by BERT as U:

$$Sim(U) = (W_{\text{STS-B}}^T \cdot U) \tag{6.2}$$

W are the weights specific to STS-B, and U is the concatenation of a dialogue context or fact with a target response.

Pair-wise text classification is used for the three relevant tasks of RTE, QQP, and MRPC. It functions in the same manner as single-sentence classification, with one difference. Two, instead of one, sequences are used as input to the model. The dialogue context or fact and the target response are concatenated. Between the two, a special token is inserted to signify that the input sequence has two components:

$$\begin{gathered} P_r(c_{task}|U) = \text{softmax}(W_{task}^T \cdot U), \\ task \in \{\text{RTE}, \text{QQP}, \text{MRPC}\} \end{gathered} \tag{6.3}$$

Pairwise ranking finds its application in the QNLI task. Likewise to pairwise text similarity, The dialogue context or fact and the target response are concatenated C (or F) R from the same dialogue are encoded as one U to calculate a relevance score:

$$\begin{aligned} Rel(U) &= g(W_{QNLI}^T \cdot U), \\ g(x) &= \frac{e^x}{e^x + 1} \end{aligned} \tag{6.4}$$

After model predictions are made on all utterances and sequential pairs of those across all tasks, the outputs have been rescaled between 0 and 1 for each GLUE task independently, as well as the scores given by the human annotators.

$$x'_{TASK} = \frac{x_{TASK} - min(x_{TASK})}{max(x_{TASK}) - min(x_{TASK})} \tag{6.5}$$

Finally, similarly as Mehri & Eskenazi [57], we train a regression that combines all the scores in one overall score:

$$y_{overall\_score} = b + \sum_{i=0}^{GLUE} w_i \cdot x_i \tag{6.6}$$

6.4 Evaluation

Here, we analyze the dialogue datasets [57] for possible relations between the GLUE task predictions and the annotator scores.

6.4.1 Baseline: UnSupervised and Reference free (USR) evaluation metric

To bring the results into context, we compare our results to the work of Mehri & Eskenazi [57]. Their approach is reference-free and unsupervised. So, it acts as a baseline against which we compare the method proposed in this chapter. The algorithm has three components.

The first component, RoBERTa [32], is fine-tuned on either PersonaChat [85] or Topical-Chat [93]. A concatenation of the input dialogue context and the target response is provided to its masked language modelling (MLM) objective. The tokens in the response part are iteratively replaced. In the end, the approach provides a probability score for the whole target sequence that indicates its fluency given the dialogue context. It is referred to as *USR-MLM*.

The second component again uses RoBERTa as its foundation. However, this time, it is fine-tuned on the Ubuntu Corpus [94] to perform dialogue retrieval using negative sampling. It is trained to distinguish between the proper response of a given context and a randomly sampled one. Mehri & Eskenazi [57] report that this metric is appropriate for evaluating Maintains Context, Interesting, and, Uses Knowledge. They refer to it as *USR-MLM* $(x = c)$ or *USR-MLM* $(x = f)$ for calculating it against the dialogue context or dialogue facts, respectively.

Finally, the third component is a combination of the other two. Mehri & Eskenazi [57] propose using a regression model to obtain one single score based on two separate metrics. This enables measuring the overall quality of a conversation. It is referred to as only *USR*.

While Mehri & Eskenazi [57] report turn- and system-level correlation scores. We benchmark only against turn-level scores due to a lack of detail of how the system-level ones are calculated.

6.4.2 Quantitative Assessment

In Tables 6.1(a) and 6.1(b), we present the correlation analysis between the automated quality metrics and human annotator scores.

In almost all of the criteria, the combined proxy indicators via linear regression outperform the combined USR metric and its best-performing components. Whereas, in the few cases where USR performs better than the proxy indicators, it is within a minor relative difference.

Looking at the *Understandable* and *Natural* criteria, we see that CoLA as a single proxy indicator can weakly infer the two measures on the TopicalChat dataset. However, it is outperformed by STSB and MRPC in PersonaChat, which suggests that the dialogues have a different nature, that involves context more strongly. This difference is also visible in the weaker performance of USR-MLM for *Understandable* and the shift to context-based USR-DR for *Natural*.

Maintains Context is the only criterion where USR outperforms the proxy indicators. Among the proxy indicators, Semantic Textual Similarity Benchmark (STSB) is the best performer, suggesting that some partial semantic overlap between context and response is necessary to model a dialogue's cohesiveness. Although, it is common sense that a reply does not need to have a high degree of semantic overlap with its context. Ultimately, the context-based USR-DR is the best-performing measure. We contribute its performance to the fact that it has been trained on dialogue data to distinguish between a correct and randomly sampled response.

We turn our attention to the *Interesting* quality measure, where USR struggles on the PersonaChat dataset. The linear regression of the proxy indicators outperforms the rest by a considerable margin. It is curious to see that the calculated STSB against the conversation data has a relatively higher correlation score. This performance suggests that responses that used the facts from the dialogue were also considered as engaging, i.e., there is an overlap between the criteria *Interesting* and *Uses Knowledge*. Aside from that, we recommend using Recognizing Textual Entailment (RTE) to indicate the interestingness of dialogue using only its context. Our results show a weak correlation with Pearson's and Spearman's coefficients ranging from 0.11 to 0.21.

The lastly mentioned metric is also the best performer for the latter criterion. Furthermore, the fact-based STSB that is compared against *Uses Knowledge* delivers the highest correlation score among all metrics. Thus, a kind of semantic similarity measure can be very indicative of whether a knowledge base is mentioned in a conversation or not.

The linear regression of all proxy indicators appears as the most consistent performer delivering the highest scores among several specific criteria and for *Overall* one except for the context-based USR-DR, which has a higher Spearman correlation score.

All of the correlation coefficients for all pairs of predictors and human annotator criteria are available in Appendix A.2.

TopicalChat		
Metric	**Pearson**	**Spearman**
Understandable		
USR-MLM	0.3268	0.3264
USR	0.3152	0.2932
CoLA	0.2458	0.2341
Lin-Reg (all)	**0.3420**	**0.3390**
Natural		
USR-MLM	0.3254	0.3370
USR	0.3037	0.2763
CoLA	0.2069	0.1677
Lin-Reg (all)	**0.3357**	**0.3130**
Maintains Context		
USR-DR (x=c)	0.3650	0.3391
USR	**0.3769**	**0.4160**
STSB	0.2350	0.2340
Lin-Reg (all)	0.3489	0.3409
Interesting		
USR-DR (x=c)	0.4877	0.3533
USR	0.4645	0.4555
STSB (fact)	0.4147	0.4103
Lin-Reg (all)	**0.5335**	**0.5364**
Uses Knowledge		
USR-DR (x=f)	0.4468	0.2220
USR	0.3353	0.3175
STSB (fact)	0.4808	0.4522
Lin-Reg (all)	**0.5119**	**0.5295**
Overall		
USR-DR (x=c)	0.3245	0.4068
USR	0.4192	0.4220
STSB (fact)	0.3324	0.3220
Lin-Reg (all)	**0.4974**	**0.4877**

(a)

PersonaChat		
Metric	**Pearson**	**Spearman**
Understandable		
USR-MLM	0.1186	0.1313
USR	**0.1324**	**0.1241**
STSB	0.1286	0.1159
Lin-Reg (all)	0.1214	0.1218
Natural		
USR-DR (x=c)	0.2291	0.1733
USR	**0.2430**	0.1862
MRPC	0.1794	**0.2410**
Lin-Reg (all)	0.1728	0.2044
Maintains Context		
USR-DR (x=c)	**0.5625**	0.6021
USR	0.5280	**0.6065**
STSB	0.3620	0.3463
Lin-Reg (all)	0.4029	0.3707
Interesting		
USR-DR (x=c)	0.2634	0.0606
USR	0.0171	0.0315
STSB (fact)	**0.3419**	**0.3378**
Lin-Reg (all)	0.3272	0.3306
Uses Knowledge		
USR-DR (x=c)	0.6309	0.4508
USR	0.3177	0.4027
STSB (fact)	**0.7329**	**0.7173**
Lin-Reg (all)	0.5921	0.5898
Overall		
USR-DR (x=c)	0.4814	**0.6087**
USR	0.4693	0.4115
STSB (fact)	0.3742	0.3898
Lin-Reg (all)	**0.5290**	0.5382

(b)

Table 6.1: Turn-level correlation results based on the sample dialogues from the TopicalChat (a) and PersonaChat (b) datasets. The USR metrics are from the original work of Mehri et al [57]. Only the best performing metrics are shown in the table. All of the correlation coefficients are with a statistical significance of $p < 0.05$.

6.4.3 Ablation Study

We investigate four configurations for using a different subset of the proxy indicators to calculate a combined score using linear regression and check the correlation coefficients against the various dialogue criteria:

- **Lin-Reg (single)** - a linear regression combining only the single-sentence GLUE tasks applied on the target response - CoLA, SST-2
- **Lin-Reg (context)** - a linear regression combining only the pair-wise sentence GLUE tasks that model the dialogue context, and the target response - MRPC, QQP, STSB, QNLI, RTE
- **Lin-Reg (fact)** - a linear regression combining only the pair-wise sentence GLUE tasks that model the dialogue facts, and the target response - MRPC, QQP, STSB, QNLI, RTE
- **Lin-Reg (all)** - a linear regression combining all of GLUE tasks that model the dialogue context, fact, and the target response

The combination of single sentence tasks shows signs of capability only on the criteria which can be evaluated utterance-wise, *Understandable*, *Natural*, and *Interesting*. While in the others, there is a drop in correlation coefficients and statistical significance, which agrees with general intuition. The single-sentence tasks cannot model dialogue quality metrics that require a view beyond the single utterances.

Turning to *Maintains Context*, we see the inverse perspective. The pair-wise sentence proxy indicators applied to the dialogue context, and target response demonstrate the best ability, while the single sentence is the worst. Furthermore, the observation is partially supported by the pair-wise tasks applied to the dialogue facts.

In regards to *Interesting*, it is evident that the pair-wise tasks outperform the single-sentence ones since context dictates what is engaging in a conversation rather than the single utterances.

Moreover, the fact-based pair-wise proxy indicators demonstrate their strong ability to model the *Uses Knowledge* criterion since these are the only automatic metrics that have access to the fact information. In comparison, the others underperform since they are not evaluated against the relevant data.

Finally, it is evident that to calculate an *Overall* score, one needs to use all of the proxy indicators. All of the subset combinations perform worse than the linear regression combining all of the metrics. Moreover, we see how the correlation improves for the combined score regarding the specific criteria like *Maintains Context*, and *Interesting*.

6.4.4 GLUE Predictor Feature Importance

In Figure 6.2, we present the inferred weights of the single GLUE predictors via linear regression.

It is immediately evident that in both datasets, the single sentence tasks, CoLA and SST, have an insignificant influence on the prediction of the overall quality score.

Semantic overlap between the utterances via STS-B and MRPC plays in both cases a significant role. However, in TopicalChat, the latter of the two has an even more substantial part. The trivia-like nature of the conversations explains the behavior. The significant scores of QQP and QNLI between facts and conversation utterances support the observation.

TopicalChat		
Metric	**Pearson**	**Spearman**
Understandable		
Lin-Reg (single)	0.2542	0.2470
Lin-Reg (context)	0.1664	0.1638
Lin-Reg (fact)	0.2572	0.2362
Lin-Reg (all)	0.3420	0.3390
Natural		
Lin-Reg (single)	0.2148	0.1853
Lin-Reg (context)	0.1986	0.1972
Lin-Reg (fact)	0.2244	0.1805
Lin-Reg (all)	0.3357	0.3130
Maintains Context		
Lin-Reg (single)	*0.0469*	*0.0197*
Lin-Reg (context)	0.2859	0.2946
Lin-Reg (fact)	0.2272	0.1921
Lin-Reg (all)	0.3489	0.3409
Interesting		
Lin-Reg (single)	0.1483	*0.0881*
Lin-Reg (context)	0.3884	0.4008
Lin-Reg (fact)	0.4358	0.4078
Lin-Reg (all)	0.5335	0.5364
Uses Knowledge		
Lin-Reg (single)	0.0699	0.0377
Lin-Reg (context)	0.2455	0.2751
Lin-Reg (fact)	0.5517	0.5182
Lin-Reg (all)	0.5119	0.5295
Overall		
Lin-Reg (single)	0.1432	0.1138
Lin-Reg (context)	0.3492	0.3587
Lin-Reg (fact)	0.3897	0.3482
Lin-Reg (all)	0.4974	0.4877

(a)

PersonaChat		
Metric	**Pearson**	**Spearman**
Understandable		
Lin-Reg (single)	*0.0643*	*0.0603*
Lin-Reg (context)	0.1626	0.1345
Lin-Reg (fact)	*0.0255*	*0.0328*
Lin-Reg (all)	0.1214	0.1218
Natural		
Lin-Reg (single)	*-0.0285*	*0.0302*
Lin-Reg (context)	0.2033	0.2160
Lin-Reg (fact)	*0.0546*	*0.0319*
Lin-Reg (all)	0.1728	0.2044
Maintains Context		
Lin-Reg (single)	*0.0974*	*0.1012*
Lin-Reg (context)	0.4178	0.3981
Lin-Reg (fact)	0.1783	0.1110
Lin-Reg (all)	0.4029	0.3707
Interesting		
Lin-Reg (single)	0.1675	0.1597
Lin-Reg (context)	0.2185	0.2216
Lin-Reg (fact)	0.3446	0.3412
Lin-Reg (all)	0.3272	0.3306
Uses Knowledge		
Lin-Reg (single)	*0.0464*	*0.0644*
Lin-Reg (context)	0.1909	0.1916
Lin-Reg (fact)	0.6959	0.7020
Lin-Reg (all)	0.5921	0.5898
Overall		
Lin-Reg (single)	0.1216	0.1263
Lin-Reg (context)	0.3975	0.3802
Lin-Reg (fact)	0.3990	0.4135
Lin-Reg (all)	0.5290	0.5382

(b)

Table 6.2: Turn-level correlation results for different mixtures of proxy indicators based on the sample dialogues from the TopicalChat (a) and PersonaChat (b) datasets. All of the correlation coefficients except the ones with *italics* have a statistical significance of $p < 0.05$.

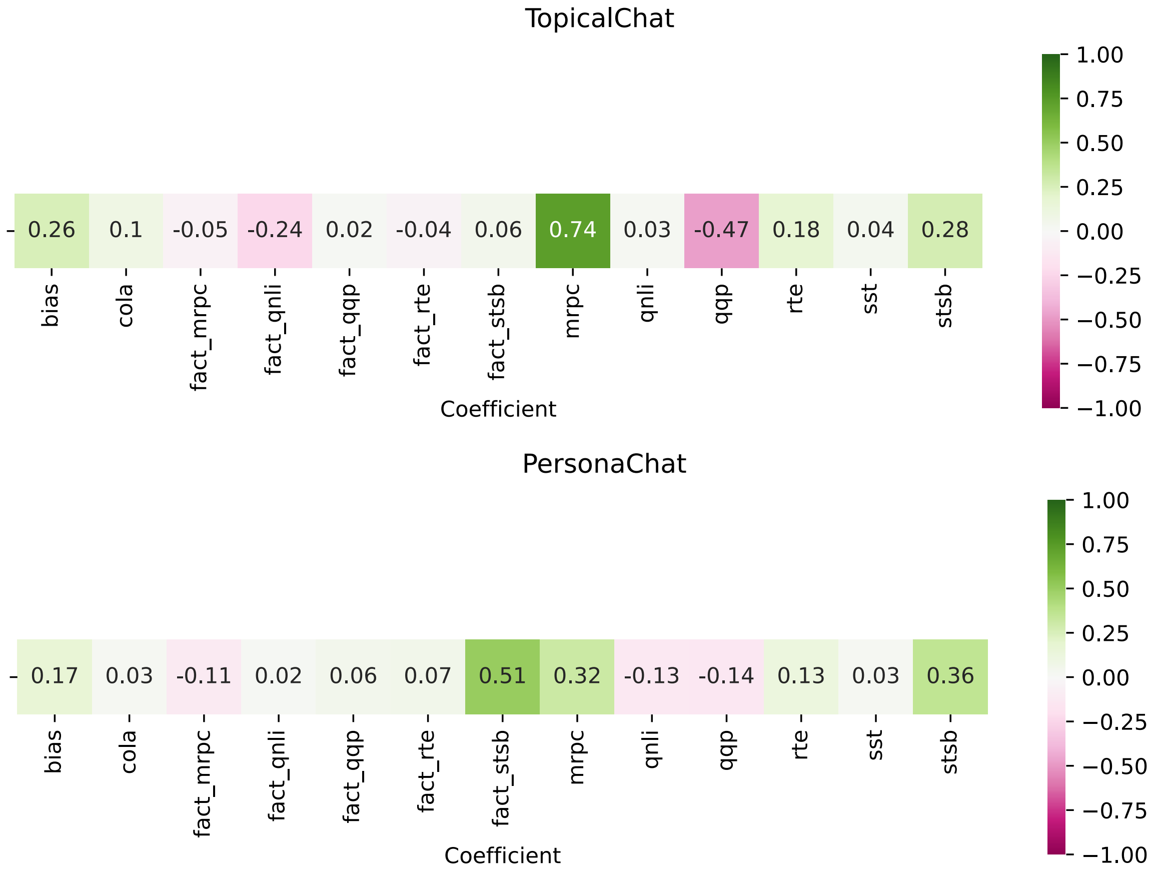


Figure 6.2: The weights as inferred by the linear regression **Lin-Reg (all)**) for each of the single GLUE predictors.

Looking at the influence of knowledge-base-related predictors, we see that in PersonaChat, it is essential to have semantic similarity (STSB) with the knowledge base facts, i.e., that the dialogue systems use the personal traits in the conversation.

6.4.5 Error Analysis

We provide regression plots with 95% confidence intervals between predictions and human annotator scores. Figures 6.3, and 6.4 show the correlation between the single GLUE predictions and the human annotator scores for TopicalChat and PersonaChat, respectively. While, Figures 6.5, and 6.6 show the correlation between the various combinations using linear regression and the human annotator scores for TopicalChat and PersonaChat, respectively. The vertical lines represent the prediction distribution for the given averaged annotator score within a 95% confidence interval. The dot signifies the mean value. For example, looking at Figure 6.6, subplot "lin-reg_fact | Uses Knowledge," the line overlaps well with the lowest (0) and the highest score (1), meaning that the prediction can distinguish well between when a dialogue uses knowledge or not. However, in the cases where the annotators could not agree, the predictor tends to overestimate them using knowledge since the intervals are below the regression line.

Based on the plots, we draw the following key conclusions:

- The linear regression on all scores has a decent general performance. Its weakness is the lower-end spectrum of the human-annotator overall quality criteria. There is a higher score variance, i.e., higher disagreement between the annotators.
- STS-B performs well on the "clear-cut" samples where knowledge is used or not. However, on borderline cases, where annotators disagree, i.e., some say knowledge is used and others not, it performs worse.
- CoLA performs excellently on the samples that were marked as Understandable by all annotators. As the scores for understandability decrease, so does the inter-annotator agreement. Hence, also the performance of CoLA.

Overall, it appears that the approach suffers the most when there is a high disagreement between the annotators, which are on the lower end of the human annotator scoring.

The USR dataset includes information about the annotators in the form of nicknames. Based on those, one can assume that they were non-native English speakers with various backgrounds. Hence, there is a low inter-annotator agreement on "Understandable" and "Natural." For example, native speakers of a Romance and a Slavic language are more likely to disagree on these two criteria. Furthermore, it is also confirmed by the higher variance in the annotator score on the lower spectrum of CoLA predictions, i.e., annotators agree well, what understandable language is, but not the opposite.

6.5 Summary

This chapter considered a model trained on GLUE as a proxy indicator for the quality of knowledge-grounded dialogues offering different perspectives on dialogue quality criteria. It does not need any references or supervision and can outperform other competing approaches like USR [57]. Pearson's and Spearman's correlation coefficients suggest that single proxy indicators and their various combinations via linear regression can infer dialogue quality either on specific criteria or in general. This composable nature can be used to tune the approach to focus more on particular criteria than others.

While one might be concerned that using the approach might offer an advantage to dialogue systems incorporating BERT, we think it poses little to no risk. BERT is an encoder approach and is considered uncommon for sequence generation applications. Hence, the risk of bias is reasonably low. In addition, one could also use any other base model architecture for training GLUE predictors.

The model has no training or fine-tuning that is specifically geared towards dialogues. However, we showed that lack of exposure to conversational data could be problematic for metrics like *Maintains Context*. Hence, we set as future work to investigate additional pre-training on dialogue data similarly as Mehri & Eskenazi [57], but also considering other proxy indicators like DialogueNLI [90], which frame the natural language inference task in a conversational setting.

Finally, while we used separately trained instances of BERT for each of the GLUE tasks, one could also consider using a multi-tasking method. For example, Liu et al [95] present Multi-Task Deep Neural Networks (MT-DNN) that employ a single instance of BERT for all GLUE tasks. We believe using multi-tasking and BERT together would make its application in a productive environment much more effortless, since model weights are to a greater extent shared between the tasks.

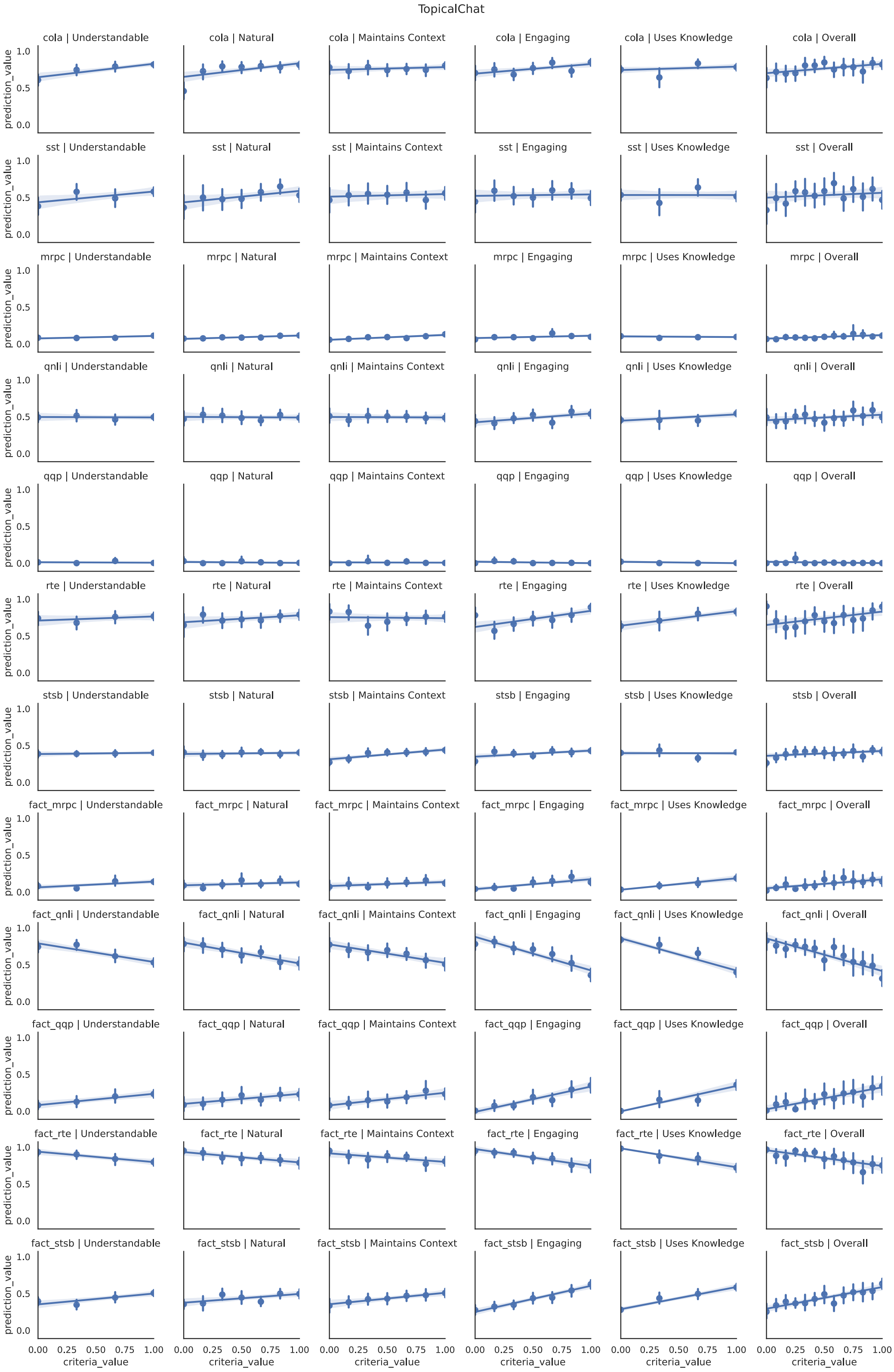


Figure 6.3: Regression plots between the single GLUE predictors and the human annotator scores on **TopicalChat**

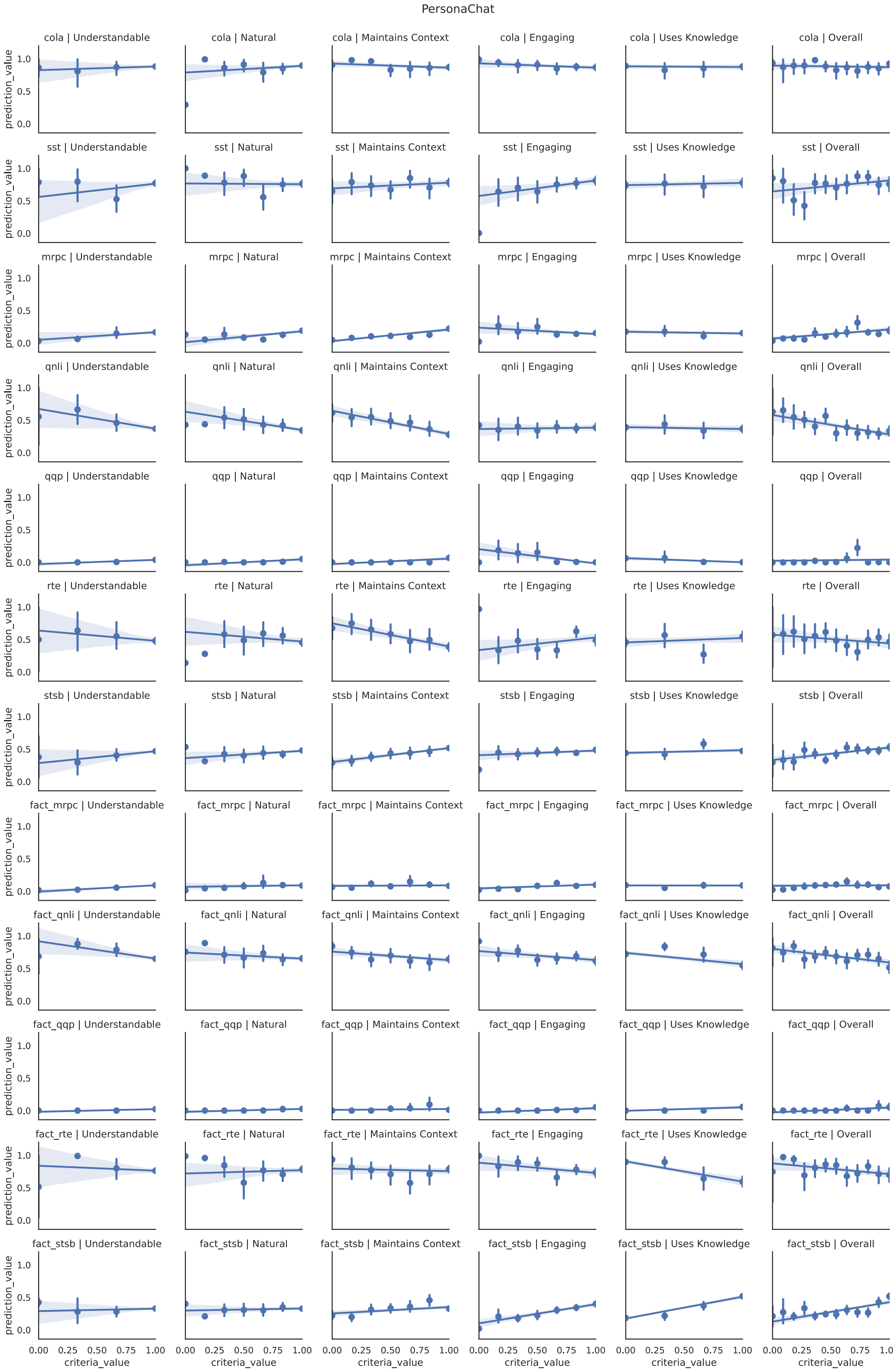


Figure 6.4: Regression plots between the single GLUE predictors and the human annotator scores on **PersonaChat**

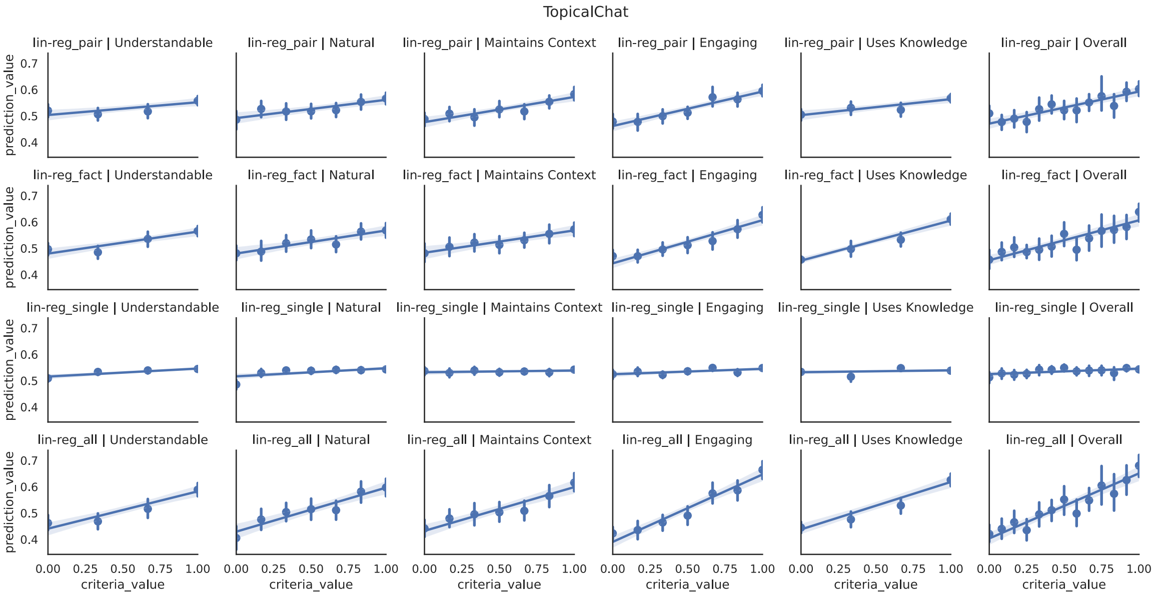


Figure 6.5: Regression plots between the combined linear regression predictors and the human annotator scores on **TopicalChat**

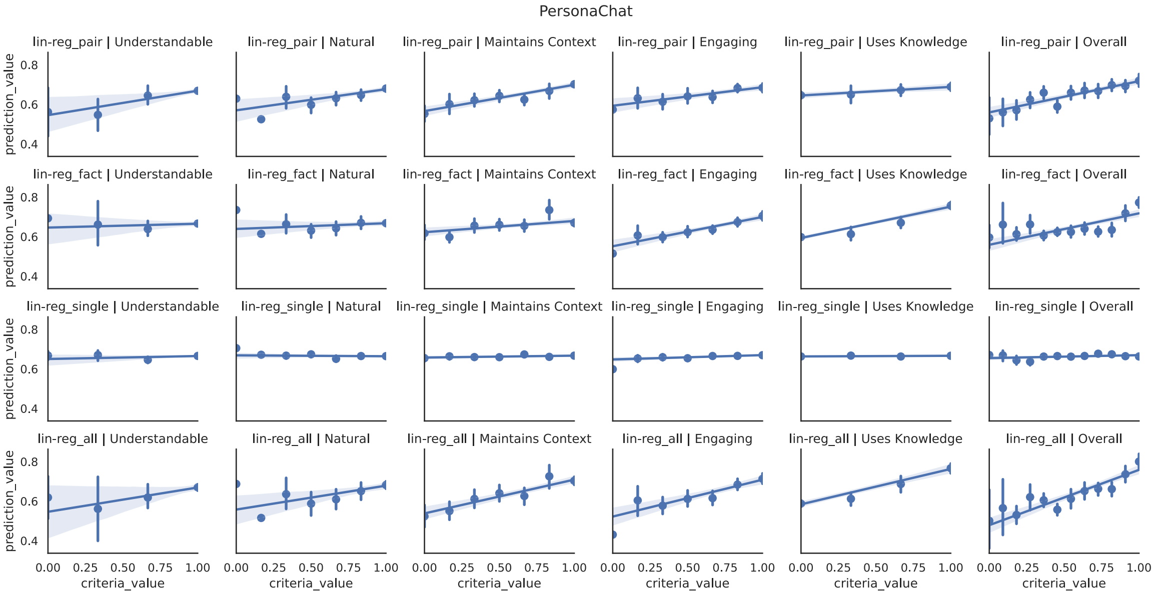


Figure 6.6: Regression plots between the combined linear regression predictors and the human annotator scores on **PersonaChat**

CHAPTER 7

Unsupervised Dialogue Breakdown Detection

In the midst of chaos, there is also opportunity

Sun Tzu

In the last three chapters, we discussed potential ways to assess quality of conversations, and consequently of dialogue systems. However, this only helps with the prevention of issues with these methods. User experience of dialogue systems can sometimes suffer from a dialogue breakdown when the system generates a reply that harms the conversation flow. Hence, the automated detection of such occurrences can prove helpful to the overall experience of dialogue systems. Their identification could help develop more natural-sounding systems and enable the triggering of a fallback strategy.

The task of dialogue breakdown detection aims to benchmark an algorithm's ability to detect responses that could adversely affect the conversation flow and cause a dialogue breakdown. The Dialogue Breakdown Detection Challenge 4 (DBDC4) [7] is designed to evaluate systems how well they are identifying such dialogue breakdowns. The shared task has been designed with three possible classes in mind: *Not a Breakdown (NB)*, *Possible Breakdown (PB)*, and *Breakdown (B)*. Multiple annotators have labeled the utterances within a conversation. We will revisit the task and its dataset in greater detail in Section 7.3.1.

Stepping into the the shoes of a human annotator yet again, a person can notice a dialogue breakdown through contrast, i.e., the conversation feels initially "odd" compared to what we experience as conversations in everyday life. Humans have a good intuition when a dialogue is "usual", so when they face a dialogue generated by a machine-learning algorithm, it can potentially feel "out of distribution."

Research Question 4

Do out-of-distribution detection methods detect breakdown in a conversation?

Earlier work has employed feature-engineered algorithms like random forests [64], LSTM-based neural networks [63] and even pre-trained language models like BERT [62] or XLM [65]. However, all of them have one common property. All of these approaches employ supervised models that use labeled data for training.

This chapter uses the following article as a foundation:

> **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2022. An Unsupervised Baseline For Dialogue Breakdown Detection Using Ouf-of-distribution Detection Methods. In Review for the 26th International Conference on Artificial Intelligence and Statistics (**AISTATS**).

In this chapter[1], we propose an unsupervised approach that relies only upon dialogue data pre-training. To the best of our knowledge, this is the first approach for solving the task without any direct supervision. In our proposal, we combine a pre-trained generative language model, GPT2 [3, 91], with out-of-distribution approaches to tackle the problem in an unsupervised manner. Our experiments show that the system has comparable performance to models that use supervision from the DBDC4 data. To our knowledge, it is also the only unsupervised approach on the benchmark that is publicly known.

7.1 Background

Larson et al. [80] suggest outlier identification for data annotation in dialog datasets to identify incorrect utterances. Their method averages word embeddings of a reply's text to achieve an utterance level representation. The next stage clusters the vectors. It then calculates the distances for each utterance from the center (mean) of the group, and ones furthest away are deemed abnormal. The methodology offers little details on the coherence of the discussion. It does not provide a substitute or even support for human annotators. Until very lately, Sai et al [55], and Mehri et al. [57] propose the usage of language models as indicator of dialogue quality. All of these approaches require no references or supervision.

Out-of-distribution detection is also familiar under other names like anomaly or outlier detection. Depending on the context, it might have slight variations in meaning or problem definition. In the context of this chapter, we will work with the following general description for out-of-distribution detection. A machine learning model is trained on in-distribution data and is requested to make predictions on both in- and out-of-distribution samples. In-distribution test samples are from the same distribution as the training results. It is assumed that the trained neural network will consistently work on examples that follow same or similar patterns as the training data.

On the other hand, samples that do not correspond to the distribution are anomalous observations. Predictions based on those are considered less reliable. The neural network does not know such data and is not able to model it. Out-of-distribution detection aims to recognize such samples [39].

In the dialogue breakdown detection context, we can assume that a deep-learning-based dialogue system is trained on a dataset (its in-distribution samples). While performing inference, the generated response might potentially be erroneous, which is by no means comparable to what is seen in the original dataset. We want to treat those faulty samples as out-of-distribution. Thus, we can detect if and to what degree the utterance can lead to a dialogue breakdown.

[1] Complete code and resources to reproduce the work can be found on the following link: https://doi.org/10.60507/FK2/MAVB6H

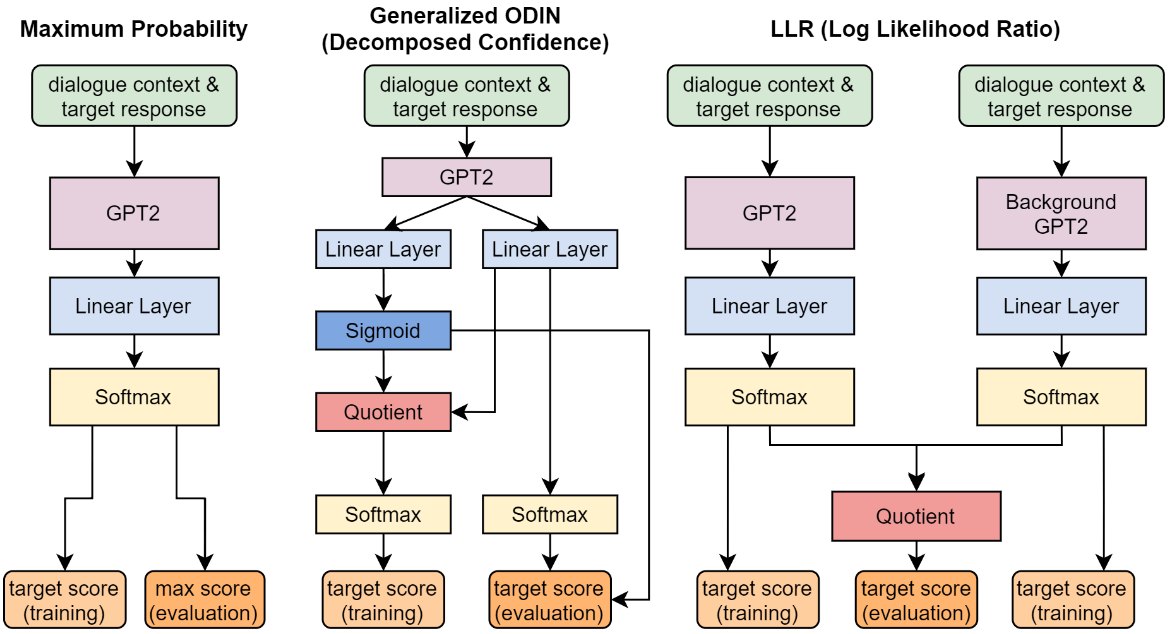


Figure 7.1: An overview of the architectures of the three OoD approaches: Maximum Probability [40], Decomposed Confidence [42], Likelihood Ratios [43].

We have discussed OoD detection in greater detail in Section 2.4.

7.2 Approach

We have picked the following list of approaches based on previous works. Note, some of them are not related to natural language processing:

1. According to Hendrycks and Gimpel [40], *softmax-based classifiers* are likely to have lower scores when they have OoD inputs compared to regular ones.

2. Hsu et al. [42] propose a modified version of ODIN [41], *Decomposed Confidence* or *Generalized ODIN*, that does not need tuning on OoD samples, which are hard to define and acquire. Thus, they use randomly perturbed input instead.

3. Ren et al. [43] suggest *likelihood ratios* for estimating whether a sample is in- or out-of-distribution. They apply two instances of the same model. The first one is trained as usual on in-sample data. The other one is trained on randomly perturbed data. The ratio of probability scores of the two indicates whether a sample is OoD.

We proceed with a detailed description of the approaches mentioned above for out-of-distribution detection and show how we apply them for evaluating open-domain dialogues in greater detail. We present an overview of the three architectures in Figure 7.1.

7.2.1 Maximum Class Probability

Hendrycks and Gimpel [40] define a framework of tasks in computer vision, natural language processing, and automatic speech recognition to investigate the behavior of softmax classifiers depending on the input data. They show that correctly classified examples have higher probability scores when compared to faulty classifications. Thus, they make a series of experiments where they look into the maximum scores of softmax outputs. Hendrycks and Gimpel show that the classifier consistently outputs lower maximums scores for OoD samples. In contrast, in-distribution samples have higher scores.

Dialogue systems commonly use a softmax distribution over a set of possible tokens (vocabulary) to generate a response. Similarly, we will use the maximum class probabilities from each position in an utterance to obtain a sort of quality score on token level:

$$\begin{aligned} &P_{OoD}(u_2|u_1) \\ &= \\ &P(t_{21}, t_{22}, ..., t_{2n}|t_{11}, t_{12}, ..., t_{1m}) \end{aligned} \tag{7.1}$$

where u_1 and u_2 are two consecutive utterances in a dialogue, and their respective tokens, $(t_{11}, t_{12}, ..., t_{1m})$ and $(t_{21}, t_{22}, ..., t_{2n})$. The higher the probability is, the less likely a dialogue breakdown is. In other words, P_{OoD} predicts a probability for *Not A Breakdown (NB)*.

7.2.2 Generalized ODIN: Decomposed Confidence

Liang et al. [41] propose ODIN (Out-of-Distribution detector for Neural networks). At its core, the method consists of two components. The first one is temperature scaling that is applied on softmax output:

$$S_i(x;T) = \frac{exp(f_i(x)/T)}{\sum_{j=1}^{N} exp(f_i(x)/T)}, \tag{7.2}$$

for the i-th output class, x is the input, and T is the temperature scaling coefficient. The second component is the application of random perturbations to the input:

$$\widetilde{x} = x - \epsilon \cdot sign(-\nabla_x \log S(x;T)), \tag{7.3}$$

where ϵ is perturbation magnitude. The authors report that this further pre-processing increases the softmax score gap between the in- and out-of-distribution data.

However, Hsu et al. [42] criticize the fact that both T and ϵ require OoD data to be tuned, which in certain use cases is not available. Hence, they discuss a decomposed confidence that consists of the joint class-domain probability and the domain probability:

$$P(y|d_{in}, x) = \frac{P(y, d_{in}|x)}{P(d_{in}|x)}, \tag{7.4}$$

x being the input and, y is the output class. To model those probabilities, they propose the following implementation for training:

DBDC4 English Eval	
# sessions	200
# instances	2000
# utterances per dialogue	11+

Table 7.1: Key numbers of the DBDC4 evaluation dataset.

$$P(y|d_{in}, x) = f_i(x); \tag{7.5}$$

$$f_i(x) = \frac{h_i(x)}{g(x)}; \tag{7.6}$$

$$g(x) = \sigma(w_g f^p(x) + b_g); \tag{7.7}$$

$$h_i(x) = w_i^T f^p(x) + b_i, \tag{7.8}$$

where $f_i(x)$ is the logit for the i-th class, $f^p(x)$ is the output of the penultimate layer of the network after applying the input, x, w and b are trainable parameters. For performing out-of-distribution detection inference, they suggest using $S_{DeConf} = max_i\ h_i\ or\ g(x)$. We report results using both.

Similarly as in 7.2.1, we will use Equation 7.1 to obtain a score from the conversation context and its following response.

7.2.3 Likelihood Ratios

In their work on image and genome sequence classification, Ren et al. [43] follow a similar intuition like Hsu et al. [42], where the OoD detector makes use of two components: a background component and a semantic component. The former models the population as a whole, whereas the latter is capturing patterns related to the domain data.

The background model is trained on perturbed in-domain data. Ren et al., [43] report using an independent and identical Bernouilli distribution with a rate of μ to decide which characters to be replaced with a random one. They report that $\mu \in [0.1, 0.2]$ achieves good performance for most of their experiments. Inspired by BERT's [2] masked language modeling (MLM) objective, we set $\mu = 0.15$ to decide which words from an utterance will be replaced with random other ones.

The log-likelihood ratio (LLR), i.e., the out-of-distribution detection score, is computed by using the probability scores from the background and semantic model:

$$LLR(x) = \log \frac{P_\theta(x_n|x_{<n})}{P_{\theta_0}(x_n|x_{<n})}, \tag{7.9}$$

where P_θ and P_{θ_0} are the softmax probabilities from semantic and background models, respectively. x_n is the n-th token, preceeded by the $x_{<n}$ tokens. x represents the concatenations of context utterances and response.

7.3 Setup

In this section, we report details on the setup used for conducting experiments. We describe the used datasets for evaluation and pre-training, the core model that we extended for out-of-distribution detection, and finally, how we calculate the relevant scores and classification predictions.

7.3.1 Dialogue Datasets

We evaluate our approach against the data from the Dialogue Breakdown Detection Challenge 4 [7]. Initially, the shared task has two tracks for Japanese and English languages. In our work, we focus only on English. Furthermore, we work only with the evaluation data since our approach is unsupervised and requires no training on the specialized data.

The dialogues were generated from sessions using the IRIS [96] dialogue system and the Conversational Intelligence Challenge 2 (ConvAI2) [79] datasets. Fifteen in total annotators have labeled each of the instances with one of the following labels: *Not a Breakdown (NB)*, *Possible Breakdown (PB)*, and *Breakdown (B)*.

For further details regarding the DBDC4 dataset, we kindly forward the reader to the original work of Higashinaka et al. [7].

7.3.2 GPT2

To evaluate the aforementioned out-of-distribution detection approaches, we use DistilGPT2 [91] as the foundation for conducting experiments. It is based on OpenAI GPT2 [3], which uses the now common transformer-based language model [15]. DistilGPT2 is a "condensed" version of the original GPT2 obtained by employing knowledge distillation from a bigger twelve-layer model. It results in a smaller six-layer version of the neural network, i.e., twice as few parameters, but very similar language modeling performance.

Originally, GPT2 was trained on non-dialogue data. Hence, it requires tuning on conversational data. To train the OoD detectors, we use Reddit comment chains collected for the Dialog System Technology Challenge (DSTC) 8 [97] that were written in October 2017 and November 2018. Also, in terms of their nature, they have a perfect fit for open-domain dialogues. The data has been filtered in various ways to improve quality and restrict offensive language use. The final training dataset contains 5,085,113 conversations.

However, we train for only three epochs of random subsets of the dialogues with five percent due to limited computational resources. However, our results show no difference between training for one or more epochs. In total, we train three different models for the different OoD approaches. Depending on the model, an epoch takes between eight and twelve hours. We used an Nvidia Geforce RTX 3080 GPU.

7.3.3 Benchmark

Originally, DBDC4 [7] proposes two sets of evaluation metrics. The first group treats the benchmark as a classification problem using the following three metrics:

- **Accuracy**: the number of properly detected utterances in one of the three classes (NB, PB, B) divided by the total number of examples.

- **F1 (B)**: F-measure where there two possible outcomes (binary classification):
  - Breakdown (B)
  - Possible Breakdown (PB) and Not a Breakdown (NB)
- **F1 (B+PB)**: F-measure where there two possible outcomes (binary classification):
  - Breakdown (B) and Possible Breakdown (PB
  - Not a Breakdown (NB)

In addition, the benchmark uses distribution/based metrics, Jensen-Shannon divergence (JSD) and Mean Squared Error (MSE) that also use various class configurations. They compare the predicted distribution with the one done by the fifteen annotators:

- **JSD (NB, PB, B)**: the Jensen-Shannon divergence measured over the three classes.
- **JSD (NB, PB+B)**: the Jensen-Shannon divergence measured over the two classes like in F1 (B+PB).
- **JSD (NB+PB, PB)**: the Jensen-Shannon divergence measured over the two classes like in F1 (B).
- **MSE (NB, PB, B)**, **MSE (NB, PB+B)**, **MSE (NB+PB, B)**: Mean Squared Error used in the same configurations as Jensen-Shannon

However, due to the inherent nature of OoD detection approaches, we used only the metrics that model the problem as binary classification, i.e. we do not report on accuracy, JSD (NB, PB, B), and MSE (NB, PB, B).

We report the scores of the following works to benchmark against our proposal:

- **Baseline** [7]: A Baseline approach proposed by the organizers of DBDC4. It takes advantage of conditional random fields (CRF) with extracted textual features from the dialogue context and response.
- **BitTalk** [63]: A bi-LSTM combined with self-attention.
- **NTT** [62]: A BERT-based classifier that uses conversation history, target answer, but also additional textual features.
- **RSL** [64]: An ensemble of models that employ random forests regression model and an LSTM on manually hand-crafted features.
- **CXM** [65]: Co-attentive Cross-Lingual Neural Model is based on XLM-R [66], which is a variant of RoBERTa [32] pre-trained with cross-lingual data. In addition, it also employs co-attentive encoding to better model the interaction between dialogue history and target response. The model is applied in two configurations, single language (CXM-S) and cross-lingual (CXM-D), with two languages.

Model	**F1**	**F1**	**JSD**	**JSD**	**MSE**	**MSE**
	(B)	**(B+PB)**	**(NB, PB+B)**	**(NB+PB, B)**	**(NB, PB+B)**	**(NB+PB, B)**
Baseline	0.3421	0.5803	0.3176	0.2670	0.2788	0.2295
BitTalk	0.3901	0.6492	0.0706	0.0570	0.0734	0.0513
LIIR	0.0981	0.0984	0.4193	0.4186	0.2641	0.2853
NTT-1	0.4641	0.7664	0.0389	0.0504	0.0444	0.0465
NTT-2	0.4482	0.6369	0.0391	0.0481	0.0432	0.0429
NTT-3	0.4547	0.6724	0.0417	0.0449	0.0472	0.0399
NTT-4	0.4403	0.6079	0.0407	0.0433	0.0455	0.0384
RSL-1	0.4411	0.6740	0.0420	0.0438	0.0480	0.0398
RSL-2	0.4483	0.7276	0.0439	0.0462	0.0506	0.0414
RSL-3	0.4554	0.6961	0.0401	0.0424	0.0455	0.0381
RSL-4	0.4650	0.7174	0.0412	0.0438	0.0469	0.0389
RSL-5	0.4690	0.6947	0.0389	0.0416	0.0439	0.0369
CXM-S	0.5303	0.6471	0.0351	0.0333	0.0396	0.0318
CXM-D	0.5825	0.7756	0.0336	0.0323	0.0370	0.0315
Max-Prob	0.4149	0.6425	0.0916	0.0396	0.1566	0.0575
ODIN-h-max	0.4132	0.6434	0.0857	0.0403	0.1460	0.0584
ODIN-g	0.4105	0.6434	0.0847	0.0406	0.1440	0.0588
LLR	0.2758	0.3631	0.1825	0.0582	0.2956	0.0782

Table 7.2: Benchmark results on the DBDC4 English track. We have not included Accuracy, JSD (NB, PB, B), and MSE (NB, PB, B), because our proposal models the problem as binary classification and those measures cannot be applied. The scores of the competing approaches have been retrieved from [7], and [65]

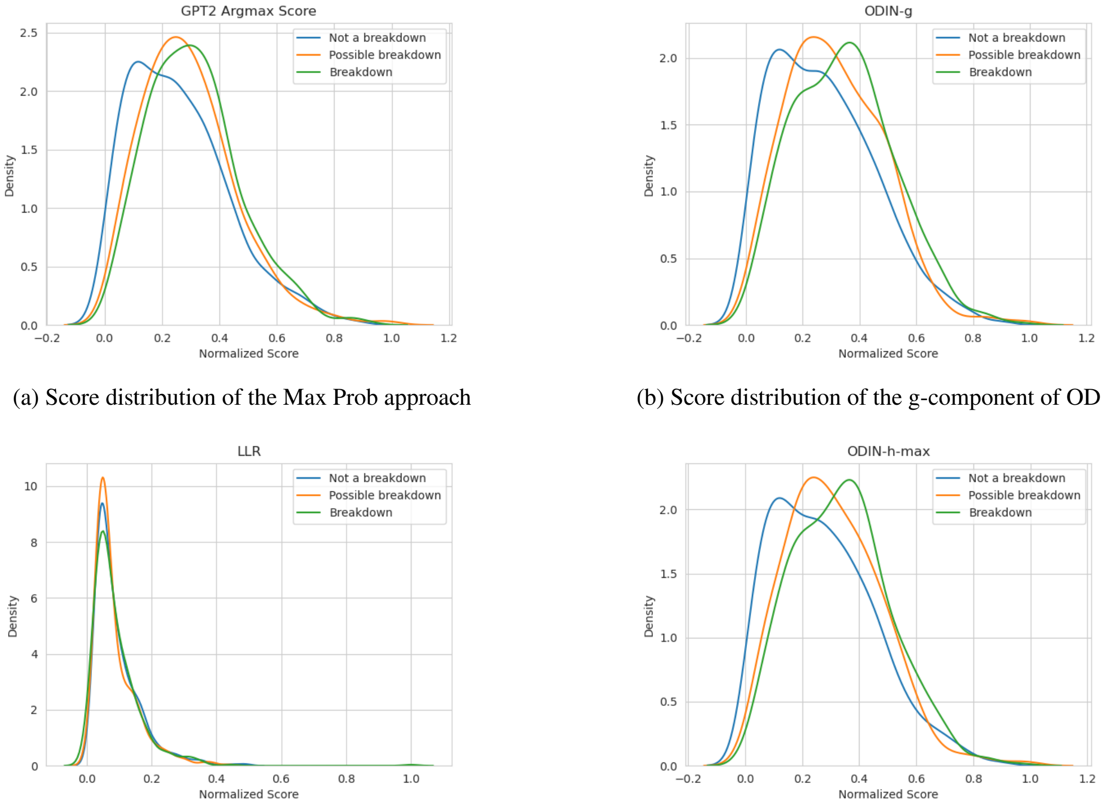
7.3.4 Scoring

To convert dialogues into a language modelling problem, all utterances within a dialogue are concatenated into one long sequence that ends with a special end-of-text token. The preceding utterances in the dialogue history are denoted as $S = x_1, ..., x_m$ and the corresponding response as $T = x_{m+1}, ..., x_n$, then the conditions probability between the two is the following:

$$P(T|S) = \prod_{i=m+1}^{n} P(x_i|x_1, ..., x_{i-1}), \tag{7.10}$$

which will then be adjusted to incorporate the necessary modifications in converting it to an out-of-distribution detector. m is the amount of tokens in the dialogue context, while n is the number of tokens in the whole concatenated sequence. Finally, after all probability outputs are obtained (including the human annotator evaluation), the scores are normalized between zero and one for each model separately:

$$\widetilde{x}_{score} = \frac{x_{score} - min(x_{score})}{max(x_{score}) - min(x_{score})} \tag{7.11}$$



(a) Score distribution of the Max Prob approach

(b) Score distribution of the g-component of ODIN

(c) Score distribution of the LLR approach

(d) Score distribution of the h-component of ODIN

Figure 7.2: Gaussian kernel density estimation of the normalized score for each of the OoD approaches. Each of the testing samples has been labeled into one of the three classes based on majority vote of the annotators.

Finally, since the out-of-distribution detection offers a binary prediction for whether the data is in (positive, one) or out (negative, zero) of the distribution, we need to inverse the score to make it fit with the problem definition of DBDC4:

$$x_{inverted\_score} = 1 - x_{score} \tag{7.12}$$

7.4 Evaluation

This section presents comparison results to previous works and performs qualitative analysis to better understand OoD-based approaches' performance.

7.4.1 Model Comparison

In Table 7.2, we show the benchmark results on the DBDC4 English track. We do not report the multi-class evaluation metrics. The inherent binary nature approaches are based on out-of-distribution detection. Some of the methods participated with more than one run in the shared task. Hence, they are listed with an integer suffix.

Immediately, it is visible that that three out of the four OoD approaches have comparable performance to the competing works, including the most current state-of-the-art system. While out-of-distribution detection does not outperform better than the other suggestions, however, the big advantage is that our proposal did not use any of the training data of DBDC4 in contrast to the other systems.

The major disadvantage of the OoD approaches is that they have worse performance when PB and B are treated as one label when compared to the supervised models. The behavior suggests that they are struggling with discriminating between cases with no breakdown or a possible one. Furthermore, we can see similar behavior with the two extremes of NB and B. Overall, the out-of-detection approaches struggle with discriminating the "middle" type of cases, where it is more challenging to make a clear distinction.

The performance of likelihood ratios (LLR) stands out as particularly bad. It fails to outperform even the original baseline of DBDC4. In the Section 7.4.2, we will attempt a qualitative analysis to understand the issue better.

7.4.2 Qualitative Analysis

In Figure 7.2, we present the distribution of the normalized scores of the four OoD approaches. Their differences in behavior become immediately apparent.

Firstly, we notice the issues with LLR. The approach is entirely unable to discriminate between a dialogue breakdown or none and all the possible degrees of them in between. Although Ren at al. [43] demonstrate LLR to work with autoregressive networks, we experience that the proposed perturbations are not as suitable for a model like GPT2, since it was initially trained with a different goal in mind. Unfortunately, the final result is a severely maimed model that cannot discriminate normal from erroneous dialogues.

We proceed with Max-Prob, which can better detect a dialogue breakdown. However, the, more or less, complete density overlap of PB and B indicates that the model is oversensitive to possible dialogue breakdowns. Unfortunately, the overlap between no breakdown and possible/certain breakdown is still too significant.

The kernel density estimates of the two ODIN components hint at what the ideal distribution should look like. For each of the three classes, there should be a peak. With no breakdown (NB) being to the leftmost side, breakdown (B) to the rightmost one, and possible breakdown (PB) between the two. Unfortunately, with ODIN, this is not quite the case, and there is a significant overlap between the three categories. However, when we compare with Max Prob, we see that ODIN has a better variance and manages to move the peak of B slightly more to the right. Furthermore, it has a smaller overlap between PB and B. These two differences show ODIN as the best performer between the OoD approaches on dialogue breakdown detection.

7.4.3 Error Analysis

We manually examined the dialogues and the various out-of-distribution scores to look for patterns in the predictions of the approaches. We made the following discoveries:

- **Short and not so engaging responses** - In few cases, the good performers (Max Class Prob, ODIN) score rather highly on not-engaging responses (e.g. -"how old are you?" -"16 years"). While being passive, i.e., not asking a question back, conversationalist is a negative feature, it is not a breakdown. In contrast, LLR gives lower probability scores, despite its worse performance.
- **Irrelevant responses to specific questions** - In situations where the target response was completely unrelated to the context (e.g. -"Congrats! I hope I'll win something too some day." -"I am so sorry to hear that."), LLR has absolutely no sensitivity to the issue. Hence, it always gives very low scores. On the other hand, the two other approaches demonstrate a behavior that is more akin to common sense.
- **Very long sentences in the context** - All three approaches seem to have difficulties modelling unusually long sentences. Whenever there is a long and difficult sentence to read, OoD tends to give high scores, although there is no dialogue breakdown.

7.5 Summary

This chapter has proposed a well-performing unsupervised model for dialogue breakdown detection, which is based on three different out-of-distribution detection methods. It performs relatively well compared to the systems that have used direct supervision, and it even outperforms a few of them. Furthermore, it is the only unsupervised approach with usable accuracy. Hence, OoD-based dialogue breakdown detection can be employed for scenarios where no labeled data is available.

Among the three OoD techniques, log-likelihood ratios, perform the worst, and maximum probability is next, closely behind generalized ODIN. We saw that the latter two could detect a dialogue breakdown. However, they cannot discriminate effectively enough. Hence, we set as future work to investigate how one could improve their sensitivity to the varying degrees of dialogue breakdown, and its minor robustness issues that we mentioned earlier.

CHAPTER 8

Evaluating Dialogue Systems via an Opinion

> A young man is embarrassed to question an older one.
>
> *Homer*

So far, we have been discussing how to evaluate dialogue systems externally via the response they provide. In other words, we ask them to provide a sample solution to a sample problem. To this day, there are two established ways to evaluate open-domain dialogue systems that are common in industry and academia alike, [79, 98]:

- Dialogue systems are asked to provide replies for specific scenarios, which are either compared against one or more possible reference responses.
- Dialogue systems are manually evaluated by a human annotator via its responses.

Both of these evaluation options are costly but also offer a by definition limited perspective.

In Chapter 5, we proposed using language models (LM) to indicate the quality of dialogues by checking their output probabilities of a given conversation. A few of the applied LMs that are also causal (GPT2, [3], XLNet, [4]) can be seen as dialogue systems in their own right since they can generate language conditioned on an input, e.g. a response to a conversation history.

Research Question 5

Can generative dialogue systems be evaluated by means of asking them whether a sample conversation is of low or high quality?

We propose[1] to use the dialogue systems themselves to evaluate open-domain dialogues. In particular, we ask already trained dialogue systems to provide opinions on a set of sample problems with their sample solutions. In other words, they shall produce output probabilities on dialogues they have not seen before and we then conduct a correlation analysis between that output and human annotator scores. The intuition behind the idea is that a well-working approach should give low probabilities for dialogues with low evaluation scores and higher when it has a positive assessment.

[1] All resources to reproduce the work are available under the following link: https://doi.org/10.60507/FK2/FX37GD.

We show that the state-of-the-art systems have higher correlations than earlier works. Hence, such a procedure can offer an additional perspective on benchmarking dialogue systems.

This chapter has its foundations in the following publication:

> **Rostislav Nedelchev**, Jens Lehmann, and Ricardo Usbeck. 2022. EDiSOn: Evaluating Dialog Systems by their Opinion on Open-domain Conversations. In Review for the 26th International Conference on Artificial Intelligence and Statistics (**AISTATS**).

8.1 Background

In their cornerstone work on natural language generation (NLG), [16] describe NLG as a pipeline consisting of six distinct stages: 1. Content Determination, 2. Text Structuring, 3. Sentences Aggregation, 4. Lexicalization, 5. Referring Expression Generation, 6. Linguistic Realization. However, with the recent rise of neural networks, such staged approaches have become unnecessary. Deep learning is capable of learning representations that can model grammatical and semantic abstractions in an end-to-end fashion, [17, 18].

There are two groups of neural network architectures that are suitable for dialogue systems. First, (causal) language models are capable of predicting the next word given a preceding sequence. Significant work by [19] demonstrates the capabilities of LSTMs to predict the next character in a sequence. Recently, language models based on transformer neural networks became popular, [2–4, 15].

Second, the encoder-decoder architecture, [13] (often referred also as sequence-to-sequence or shortly, Seq2Seq), provides a decoupling between creating a fixed length representation of the input and consequently, decoding it into a sequence. There are many works that have used this approach to develop a dialogue system, [20–23].

8.2 Approach

8.2.1 Dialogue Systems

We use a set of milestone works in the field of dialogue systems to test their "opinions." As described in Chapter 2, the set consists of two groups, namely encoder-decoder and language model architectures. We start with the former of the two.

The first model we consider is a recurrent *Seq2Seq* approach, as described by [20]. It models a dialogue as a sequence of pairs, where each pair consists of a query and a response. That is, the model considers a response as related only to the last utterance before it. The context is encoded using a recurrent neural network (RNN), and another RNN decodes the representation as the response.

Next, we use the work by [21], Hierarchical Recurrent Encoder-Decoder (*HRED*), which builds on the method of sequence-to-sequence (Seq2Seq) through considering several contextual utterances. The overall meaning of the utterances are encoded with RNN and then encoded by an additional RNN together. The rest is as stated previously in the sequence-to-sequence method.

Third, [22] propose adding a latent variable that parametrizes the context, i.e., an extended version of the HRED, Hierarchical Latent Variable Encoder-Decoder (*VHRED*). Via a two-stage process, the

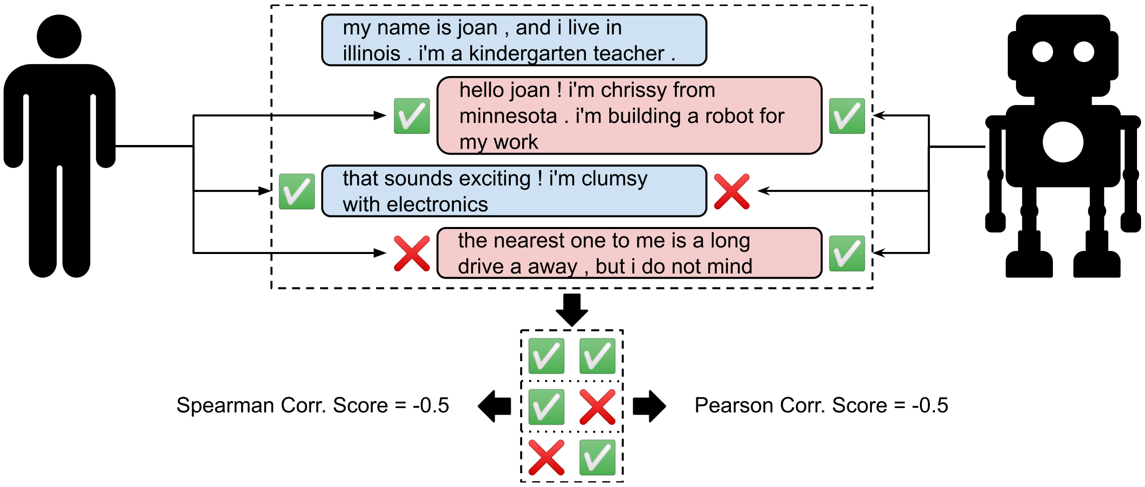

my name is joan , and i live in illinois . i'm a kindergarten teacher .
hello joan ! i'm chrissy from minnesota . i'm building a robot for my work
that sounds exciting ! i'm clumsy with electronics
the nearest one to me is a long drive a away , but i do not mind
Spearman Corr. Score = -0.5
Pearson Corr. Score = -0.5

Figure 8.1: An example demonstrating how dialogue systems are asked for an "opnion", and compared to human evaluators.

method can model hierarchically ordered sequences and generate the output sequence while retaining long-term contexts.

Finally, [23] report that VHRED has a latent variable degeneration, which makes the model almost HRED-like. They add a latent variable to the global discourse, such that it produces every word of the dialogue, instead of post-factum of the whole thread (*VHCR*).

We use the Cornell Movie-Dialogs Corpus to train the four models [83]. The dataset has 220,579 conversations and a total of 304,713 utterances. The training is executed by iteratively considering each utterance and its preceding context. The first of the four sequence-to-sequence approaches works only with a pair of utterances.

We now discuss the second group of approaches based on causal language modelling. GPT2 (Generative Pre-trained Transformer) by [3] is a standard causal LM that calculates the probability for the target token by conditioning on the series of previous ones. In the problem domain of dialogues, to use GPT2, we consider two or more consecutive utterances and capture the coherence between them. Thus, we concatenate them into one, where the source context, $S = x_1, ..., x_m$, appears first and is then followed by the target response, $T = x_{m+1}, ..., x_n$:

$$P(T|S) = \prod_{i=m+1}^{n} P(x_i|x_1, ..., x_{i-1}) \tag{8.1}$$

We use models of various sizes, [87], i.e., with 117M, 345M, and 774M parameters, that are pre-trained on English Wikipedia and various fictional works.

We also investigate DialoGPT, [8]. In essence, it uses the same architecture as GPT2. However, it has been trained on English Reddit comment threads, instead of non-dialogue data like Wikipedia. The dataset has been processed to reduce offensive language and nonsensical examples. The final data contains 147,116,725 conversations. Similarly to GPT2, we use different model sizes, i.e., 124M, 355M, and 774M parameters.

We evaluate both language models in two different settings: 1. utterance pairs, where only the last utterance before the target response is considered; 2. full context, where the whole conversation

Dataset	ConvAI1		ConvAI2	
	r	ρ	r	ρ
Seq2Seq	0.0088	0.0748	0.2754	0.3061
	(0.6820)	(0.0005)	(0.0001)	(0.0001)
HRED	0.1048	0.1234	0.2567	0.2967
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
VHRED	0.0303	0.0639	0.2803	0.2992
	(0.1601)	(0.0030)	(0.0001)	(0.0001)
VHCR	0.0284	0.0654	0.2891	0.3043
	(0.1872)	(0.0024)	(0.0001)	(0.0001)

Table 8.1: Pearson's correlation coefficients, r, and Spearman's correlation coefficients, ρ, on the two dialogue datasets' human scores and aggregated probability scores from the encoder-decoder based architectures. The numbers in parenthesis are the statistical significance scores ($p <$).

history up until the target response is used.

8.2.2 Assessed Dialogue Datasets

Like in Chapter 4, we utilize the datasets of the ConvAI1 [1, 84] and ConvAI2 [79, 85] challenges. We would like to forward the reader to Section 4.2.2 for information regarding the data.

8.2.3 Scoring

In Chapter 5.4, we reported on experimenting with various techniques to aggregate the probabilities on the dialogue level. However, one is selected due to its best results, which we also use here. We then perform a correlation analysis between $dialog\_score$ and the human annotator scores using Pearson's and Spearman's metrics.

8.3 Evaluation

We discuss the correlation results between the output probabilities ($dialog\_score$) of the approaches mentioned above and the human annotator scores. We compare earlier to more advanced approaches for dialogue modeling. Intuition dictates that more modern systems should have higher correlation with the human annotator scores.

We start with the encoder-decoder architectures, whose correlations are displayed in Table 8.1. Among the four approaches, the hierarchical encoder-decoder stands out as the best performing system on ConvAI1. The vanilla sequence-to-sequence approach performs the worst, which is expected since it has the most basic architecture compared to the others. The two variational approaches exhibit similar behavior. Thus, they cannot be distinguished using opinion-based evaluation. However, we see a performance increase from basic sequence-to-sequence through VHRED and VHCR to HRED. In contrast, all of them perform very similar on ConvAI2.

We move to the language model architectures and the correlation coefficients (Table 8.2) between output probabilities and human annotator scores. First of all, there is an immediate difference in the

Dataset	**ConvAI1**		**ConvAI2**	
	r	ρ	r	ρ
gpt2-sm	-0.0347	0.0520	0.1345	0.3055
(pair)	(0.1073)	(0.0158)	(0.0001)	(0.0001)
gpt2-sm	-0.0217	0.0773	0.1252	0.3350
(context)	(0.3132)	(0.0003)	(0.0001)	(0.0001)
gpt2-md	-0.0127	0.0559	0.1308	0.3089
(pair)	(0.5547)	(0.0095)	(0.0001)	(0.0001)
gpt2-md	-0.0222	0.0523	0.1095	0.3438
(context)	(0.3035)	(0.0152)	(0.0001)	(0.0001)
gpt2-lg	-0.0460	0.0351	0.1173	0.3106
(pair)	(0.0326)	(0.1032)	(0.0001)	(0.0001)
gpt2-lg	-0.0224	0.0631	0.1002	0.3334
(context)	(0.2986)	(0.0034)	(0.0001)	(0.0001)
dialogpt-sm	-0.0067	0.0878	0.0828	0.3426
(pair)	(0.7565)	(0.0001)	(0.0001)	(0.0001)
dialogpt-sm	0.0461	0.1274	0.2098	0.4333
(context)	(0.0323)	(0.0001)	(0.0001)	(0.0001)
dialogpt-md	0.0421	0.1308	0.0930	0.3204
(pair)	(0.0506)	(0.0001)	(0.0001)	(0.0001)
dialogpt-md	0.0253	0.1154	0.1416	0.3608
(context)	(0.2409)	(0.0001)	(0.0001)	(0.0001)
dialogpt-lg	-0.0218	0.0989	0.1597	0.3512
(pair)	(0.3122)	(0.0001)	(0.0001)	(0.0001)
dialogpt-lg	-0.0287	0.0644	0.1638	0.3947
(context)	(0.1828)	(0.0028)	(0.0001)	(0.0001)

Table 8.2: Pearson's correlation coefficients, r, and Spearman's correlation coefficients, ρ, on the two dialogue datasets' human scores and aggregated probability scores from the language-model based architectures. The numbers in parenthesis are the statistical significance scores ($p <$).

behavior between GPT2 and DialoGPT. While they have the same architecture, they have been trained on different datasets. The former on Wikipedia and fiction, while the latter on conversational texts. This difference gives DialoGPT an advantage that is visible in the comparatively higher correlations than GPT2.

GPT2 shows sensitivy towards the number of parameters in the network. The higher the number of layers, the better it performs on, i.e. the higher the correlation. DialoGPT follows mostly the same behavior with the exception of the small version on the whole context.

Furthermore, we perform an ablation study by comparing the response of the systems to changes to the input data. Intuitively, a dialogue system should perform better when presented with the complete context, and worse, when it is working with only the last utterance of the conversation history.

GPT2 is undecided between pair utterances or full context, although there is a visible difference. However, the specialized DialoGPT prefers the full context over just the last utterance before the target response.

Comparing the two groups of fundamental architectures, it is challenging to decide on a "winner". The encoder-decoder-based models seem to have the edge over the language models on the ConvAI1 dialogues. However, we observed the opposite when it comes to ConvAI2 conversations.

8.4 Summary

In this paper, we propose a new way to evaluate dialogue systems. Namely, to ask them to assess a sample solution to a sample problem instead of asking them to provide their solution to the sample problem. In other words, we ask dialogue systems to evaluate dialogues rather than to respond themselves.

We showed that a deep-learning-based system's performance depends on various factors like its architecture, number of parameters (number of layers), or the data it has been trained through calculating the correlation coefficients.

The opinion-based approach is suitable for evaluating dialogue systems. It provides an automatic metric that serves as an additional perspective on the systems' performance. Hence, it can support the research and development of such algorithms.

What the procedure is missing is a consistently structured benchmark dataset with human annotator scores. The data from the ConvAI1 and ConvAI2 challenges has a random nature and cannot guarantee it covers a comprehensive and complete spectrum of high and low-quality dialogues. Thus, we set as future work to investigate the creation of such a dataset, which will enable a consistent and even more insightful evaluation of dialogue systems.

CHAPTER 9

Conclusion and Future Work

> The unexamined life is not worth living.
>
> *Socrates*

The main research objective of this thesis is to enable automatic evaluation of dialogue systems. In Chapter 1, we laid out the research problem and examined what are the significant challenges relevant to the issue. As part of Chapter 2, we reviewed all fundamental concepts and background knowledge that serves as a basis to this thesis. Chapter 3 discussed the related work that is significant for the research goals. In the following Chapters (4, 5, 6, 7, and 8), we presented our proposals that handle the earlier defined research questions.

9.1 Review of Research Questions

To conclude this work, it is of importance to review each of the research questions and interpret the outcomes of the contributions. The core research problem was divided into five specific research questions. The first one had the goal to use autoencoders and anomaly detection for detecting low- from high-quality dialogues. The second research question demonstrated that language models can effectively infer the quality of conversations without any special training on dialogue data. The third research questions address the disadvantages of the second one by finding a component based approach that allows the measurement of separate dialogue criteria. The fourth research question aims to discover what utterances can cause a dialogue breakdown using unsupervised out-of-distribution detection methods. Finally, the fifth research question shows how dialogue systems can take the role of annotators and use this behavior to evaluate their abilities.

Research Question 1

Can anomaly detection methods be used to infer the quality of a dialogue?

Chapter 4 investigated whether anomaly detection could contrast dialogues in terms of their quality. For the purpose, we used four different dialogue modeling approaches as if they are autoencoders. Correlation scores with human annotators reveal only limited potentials. Only for a limited amount of cases the proposal could evaluate conversations. The experiments showed that the discussed neural network architectures have a limited ability to generalize, and hence, they can properly model only certain type of conversations. The issue is mostly caused due to the limited scope of the training data.

Research Question 2

Can language models indicate the quality of a conversation?

This research aims to solve the generalization weaknesses discussed in the first RQ. For the purpose, we investigated whether pre-trained language models can deduce the quality of a conversation. They have two main advantages that makes them suitable for the task. Firstly, they are trained on massive amounts of text data. This enormous exposure makes them very versatile, and there is little in terms of language usage that could "surprise" them. Second, they do not require any labelled data, but only fluent text that can have various origins.

We infer conditional probabilities using three language model approaches and dialogue data. Then, we perform a correlation analysis that demonstrates that these neural networks can evaluate conversation quality. BERT's next sentence prediction proves to be the most effective method due to its structured approach. Its then closely followed by XLNet and GPT2, which handle the task on the word level. In addition, we show that they are capable "conversationalists."

In general, all of them can be used for the task. However, the offer only a single general metric that offers no insights for the separate dialogue criteria.

Research Question 3

Are standard NLP tasks helpful with the evaluation of dialogues?

In the previous two research questions, we offered approaches that provide a metric that measures the overall dialogue quality. However, conversations are complex concepts that can be evaluated according to various criteria , e.g. fluency (i.e., how grammatical is the language use) , or coherence (i.e., how well do the responses follow the dialogue flow and topic). These dialogue features are related language skills that are specific to humans and that are also being researched as part of standardized benchmarks.

Hence in research question three, we investigated whether the tasks in the General Language Understanding Evaluation benchmark can be used to measure the quality of dialogues. For the purpose, we trained instances of the same model on the different tasks. Then, we performed inferences, collected probabilities scores related to the benchmarks, that we compared to human annotator scores on dialogue criteria in a correlation study. Our results revealed that a few of the tasks map very well to the conversation features, and can be used for automatic evaluation.

Furthermore, we demonstrated how the single evaluators can be combined together to provide a metric for overall quality. The strength of the approach is that it can use training data from other benchmarks that are not necessarily related to dialogues and still effectively infer conversation quality. In addition, the composition can be tuned to focus on different criteria which makes the metric versatile, and it can be applied on a big variety of scenarios.

Research Question 4

Do out-of-distribution detection methods detect breakdown in a conversation?

So far, in the first three research questions we discussed methods that allow to benchmark and compare dialogue systems to each other. However, another purpose to evaluate dialogue systems is to monitor them if they could run into problems where they cause a dialogue breakdown, state where an utterance breaks the conversation flow completely. Hence, we set in research question four to investigate unsupervised approaches for detecting dialogue breakdowns.

We seek inspiration in out-of-distribution detection approaches, where we adapt a standard dialogue modeling approach to the task. Among the approaches, maximum probability and ODIN show great potential in the Dialogue Breakdown Detection Challenge (DBDC) 4 benchmark. They can outperform some of the supervised approach, and according to some of the benchmark metrics, OoD-based approaches are within striking distance to the best performing method, which is also supervised.

The strength of our approach is that it is unsupervised. Hence, it requires no labelled data, but only dialogue conversations, and it can be easily applied in other languages where annotated data is scarce.

Research Question 5

Can generative dialogue systems be evaluated by means of asking them whether a sample conversation is of low or high quality?

Most of the work so far focused on evaluating dialogue systems by means of samples responses, i.e. they are presented with an example conversation, and they have to provide a response based on that history. It is they either compared to a reference answer or it is evaluated by human annotators. Ideally, we want dialogue systems to behave like humans. Hence, instead of comparing whether they "answer" like a human, we can rather check if they "evaluate" like one. So, we propose to evaluate them by their opinion.

In research question five, we investigate how we can evaluate dialogue systems by means of "opinion." We use a set of 16 different models or configurations respectively that include some of the very initial to the most advanced works. Via a correlation analysis of dialogue system's output probabilities, we demonstrate how the more recent the approach is, the better are its results. Hence, we show that dialogue modeling approaches can be evaluated by means of behavior, we do not need to generate individual response that require comparison or evaluation.

9.2 Limitations and Future Work

Future efforts that would build upon or continue the discussed works should consider the following limitations:

1. Many of the discussed methods here focus on the word or utterance level of a conversation, i.e. the provide feedback on a lower level. This requires aggregation steps (e.g. taking the average) to be taken in order to have a metric representing the whole dialogue. It is a disadvantage that definitely wastes some of the information that was captured and leads to inaccuracies. Hence, future work should focus on methods that treat a conversation as an atomic unit besides the word- and utterance-level modeling. This way, it will not be necessary to perform aggregations that loose some of the information.

2. Many of the approaches have a very strong dependency on data, especially on copious amounts, which is a disadvantage for a few reasons. In certain scenarios like languages that are spoken by significantly less people than English, it can be difficult to obtain such big datasets. Furthermore,

even if the data was somehow readily available, the training of the approaches would still require computational resources, which could be also expensive to acquire. Therefore, as a first step, future work should confirm the efficacy of the results in other languages than English that have less resources. In addition, improvements in direction of reducing the amount of necessary data should be further investigated.

3. We saw that the proxy indicators could suffer from the lack of exposure to the dialogue data. Therefore, the method could benefit from training on dialogue data, e.g. DialogueNLI [90]. In addition, the approach made use of multiple instances of BERT, which costs more computational resources. It would be beneficial to investigate a multi-tasking approach as described in the work of Liu et al. [95], where only one instance of BERT is used to tackle all the tasks in the GLUE benchmark.

4. The discussed approaches are not perfect, yet. As the various metrics (e.g. correlation or classification scores) demonstrate, there is still room for improvement. There is still a gap between human annotators the new metrics that have been discussed here. The same applies for our unsupervised dialogue breakdown detection approach.

9.3 Closing Remarks

This works investigates how dialogue evaluation can be automated to achieve parity with human annotators. Despite the challenges, we demonstrated through our experiments that machine learning and natural language processing methods can be used to estimate the quality of conversations.

In this thesis, we progressed the state of the art in referenceless, but also unsupervised, dialogue evaluation on several frontiers. In Appendix A.1, an overview with links to all resources can be found that were result of this thesis. We made the following contributions:

- We demonstrated that anomaly and out-of-distribution detection methods have potential for indicate dialogue quality.
- Our experiments showed how language models can assess conversations without training on any dialogue data.
- Our work showcased the usage of standard NLP benchmarks for dialogue evaluation, and how they can be composed together in a tuneable manner.
- We illustrated an alternative approach for evaluating dialogue systems by means of "opinion", and not "sample solution."

Future work can use our methods to further advance the state of the art in dialogue systems. The proposed methods already have practical value and can support the research and development of conversational artificial intelligence.

APPENDIX A

Appendix

A.1 Overview of Publicly Available Resources

Ch.	Title	Link
Peer-reviewed Works		
5	Language Models as Evaluators	https://github.com/SmartDataAnalytics/transformers_dialogue_evaluators
6	Proxy Indicators for Dialogue Quality	https://github.com/SmartDataAnalytics/proxy_indicators
Works in Review		
7	Unsupervised Dialogue Breakdown Detection	https://doi.org/10.60507/FK2/MAVB6H
8	Evaluating Dialogue Systems via an Opinion	https://doi.org/10.60507/FK2/FX37GD

Table A.1: Overview of the publicly available resources that are results of the work done in the thesis. The works in review are provided with anatomized links since they are still being reviewed.

A.2 Complete correlation scores for all predictors

We present complete tables with correlation scores of all pairs of predictors and human annotator scores. In Tables A.2, and A.3 are the correlation scores for the single GLUE tasks. Furthermore, Tables A.4, and A.5 present the correlation coefficients for on the GLUE predictions of the knowledge base facts and the dialogue utterances. Finally, Tables A.6, and A.7 show the correlation scores for the various combinations of the GLUE predictors using linear regression.

TopicalChat				
Predictor-Criteria	**Pearson's** r	$p <$	**Spearman's** ρ	$p <$
cola-Understandable	0.2458	0.0001	0.2341	0.0001
cola-Natural	0.2069	0.0001	0.1677	0.0014
cola-Maintains Context	0.0449	0.3959	0.0119	0.8226
cola-Engaging	0.1518	0.0039	0.0935	0.0765
cola-Uses Knowledge	0.0727	0.1686	0.0481	0.3623
cola-Overall	0.1418	0.0070	0.1136	0.0312
sst-Understandable	0.1253	0.0173	0.1114	0.0346
sst-Natural	0.1107	0.0358	0.0826	0.1176
sst-Maintains Context	0.0260	0.6225	-0.0064	0.9041
sst-Engaging	0.0146	0.7825	-0.0328	0.5346
sst-Uses Knowledge	-0.0006	0.9906	-0.0517	0.3280
sst-Overall	0.0471	0.3731	0.0139	0.7924
mrpc-Understandable	0.1216	0.0210	0.0890	0.0918
mrpc-Natural	0.1366	0.0095	0.1171	0.0264
mrpc-Maintains Context	0.2083	0.0001	0.2131	0.0001
mrpc-Engaging	0.0985	0.0619	0.0823	0.1191
mrpc-Uses Knowledge	-0.0395	0.4545	-0.0266	0.6147
mrpc-Overall	0.1419	0.0070	0.1258	0.0170
qnli-Understandable	-0.0076	0.8864	0.0062	0.9069
qnli-Natural	-0.0095	0.8571	-0.0032	0.9515
qnli-Maintains Context	-0.0078	0.8824	-0.0015	0.9768
qnli-Engaging	0.1409	0.0074	0.1538	0.0034
qnli-Uses Knowledge	0.1382	0.0086	0.1509	0.0041
qnli-Overall	0.0853	0.1060	0.0952	0.0711
qqp-Understandable	-0.0311	0.5569	-0.0369	0.4858
qqp-Natural	-0.0510	0.3346	-0.0142	0.7879
qqp-Maintains Context	-0.0173	0.7439	0.0529	0.3173
qqp-Engaging	-0.0845	0.1095	-0.0910	0.0848
qqp-Uses Knowledge	-0.1103	0.0365	-0.1352	0.0102
qqp-Overall	-0.0751	0.1548	-0.0708	0.1804
rte-Understandable	0.0598	0.2577	0.0758	0.1510
rte-Natural	0.0833	0.1147	0.0936	0.0761
rte-Maintains Context	-0.0131	0.8043	-0.0419	0.4282
rte-Engaging	0.2024	0.0001	0.2116	0.0001
rte-Uses Knowledge	0.2478	0.0001	0.2523	0.0001
rte-Overall	0.1619	0.0021	0.1554	0.0031
stsb-Understandable	0.0343	0.5160	0.0473	0.3711
stsb-Natural	0.0270	0.6094	0.0430	0.4158
stsb-Maintains Context	0.2350	0.0001	0.2340	0.0001
stsb-Engaging	0.1457	0.0056	0.1704	0.0012
stsb-Uses Knowledge	-0.0056	0.9150	0.0360	0.4962
stsb-Overall	0.1129	0.0322	0.1429	0.0066

Table A.2: Correlation scores between the GLUE tasks on the conversation utterances and the human annotator scores and their respective p-values on the **TopicalChat** dataset.

PersonaChat				
Predictor-Criteria	**Pearson's *r***	***p* <**	**Spearman's ρ**	***p* <**
cola-Understandable	0.0318	0.5828	0.0673	0.2451
cola-Natural	0.0838	0.1475	-0.0309	0.5945
cola-Maintains Context	-0.0862	0.1365	-0.1935	0.0008
cola-Engaging	-0.0665	0.2510	-0.1568	0.0065
cola-Uses Knowledge	-0.0190	0.7425	-0.1403	0.0150
cola-Overall	-0.0252	0.6635	-0.1931	0.0008
sst-Understandable	0.0743	0.1996	0.0723	0.2119
sst-Natural	-0.0064	0.9119	0.0294	0.6123
sst-Maintains Context	0.0760	0.1890	0.0988	0.0875
sst-Engaging	0.1530	0.0080	0.1242	0.0315
sst-Uses Knowledge	0.0422	0.4663	-0.0034	0.9531
sst-Overall	0.1172	0.0424	0.1068	0.0647
mrpc-Understandable	0.0857	0.1385	0.1098	0.0574
mrpc-Natural	0.1794	0.0018	0.2410	0.0001
mrpc-Maintains Context	0.3129	0.0001	0.3684	0.0001
mrpc-Engaging	-0.1266	0.0284	0.0695	0.2301
mrpc-Uses Knowledge	-0.0656	0.2574	-0.0112	0.8468
mrpc-Overall	0.1959	0.0006	0.2576	0.0001
qnli-Understandable	-0.1356	0.0188	-0.1434	0.0129
qnli-Natural	-0.1821	0.0015	-0.2058	0.0003
qnli-Maintains Context	-0.3795	0.0001	-0.3982	0.0001
qnli-Engaging	0.0163	0.7780	0.0318	0.5832
qnli-Uses Knowledge	-0.0430	0.4580	-0.0490	0.3981
qnli-Overall	-0.2553	0.0001	-0.2434	0.0001
qqp-Understandable	0.0529	0.3613	0.0830	0.1514
qqp-Natural	0.1071	0.0639	0.1857	0.0012
qqp-Maintains Context	0.1646	0.0043	0.3472	0.0001
qqp-Engaging	-0.3205	0.0001	-0.0071	0.9029
qqp-Uses Knowledge	-0.1725	0.0027	0.0208	0.7198
qqp-Overall	0.0276	0.6345	0.2125	0.0002
rte-Understandable	-0.0519	0.3704	-0.0976	0.0916
rte-Natural	-0.0710	0.2200	-0.1184	0.0404
rte-Maintains Context	-0.2789	0.0001	-0.2999	0.0001
rte-Engaging	0.1131	0.0503	0.1269	0.0280
rte-Uses Knowledge	0.0752	0.1939	0.0827	0.1531
rte-Overall	-0.0842	0.1459	-0.0766	0.1860
stsb-Understandable	0.1286	0.0259	0.1159	0.0448
stsb-Natural	0.1140	0.0486	0.1317	0.0225
stsb-Maintains Context	0.3620	0.0001	0.3463	0.0001
stsb-Engaging	0.0889	0.1242	0.0805	0.1645
stsb-Uses Knowledge	0.0988	0.0877	0.0828	0.1525
stsb-Overall	0.2591	0.0001	0.2396	0.0001

Table A.3: Correlation scores between the GLUE tasks on the conversation utterances and the human annotator scores and their respective p-values on the **PersonaChat** dataset.

TopicalChat				
Predictor-Criteria	**Pearson's r**	$p <$	**Spearman's ρ**	$p <$
fact_mrpc-Understandable	0.1357	0.0100	0.1935	0.0002
fact_mrpc-Natural	0.0564	0.2859	0.1186	0.0244
fact_mrpc-Maintains Context	0.0827	0.1174	0.1981	0.0002
fact_mrpc-Engaging	0.2017	0.0001	0.3052	0.0001
fact_mrpc-Uses Knowledge	0.3162	0.0001	0.3839	0.0001
fact_mrpc-Overall	0.1749	0.0009	0.2647	0.0001
fact_qnli-Understandable	-0.2597	0.0001	-0.2355	0.0001
fact_qnli-Natural	-0.2419	0.0001	-0.1981	0.0002
fact_qnli-Maintains Context	-0.2239	0.0001	-0.1842	0.0004
fact_qnli-Engaging	-0.4034	0.0001	-0.3727	0.0001
fact_qnli-Uses Knowledge	-0.5291	0.0001	-0.5457	0.0001
fact_qnli-Overall	-0.3784	0.0001	-0.3390	0.0001
fact_qqp-Understandable	0.1656	0.0016	0.2147	0.0001
fact_qqp-Natural	0.1217	0.0209	0.1788	0.0007
fact_qqp-Maintains Context	0.1607	0.0022	0.1917	0.0003
fact_qqp-Engaging	0.3197	0.0001	0.3824	0.0001
fact_qqp-Uses Knowledge	0.4373	0.0001	0.5350	0.0001
fact_qqp-Overall	0.2683	0.0001	0.3347	0.0001
fact_rte-Understandable	-0.1823	0.0005	-0.1896	0.0003
fact_rte-Natural	-0.1512	0.0040	-0.1408	0.0075
fact_rte-Maintains Context	-0.1297	0.0138	-0.1398	0.0079
fact_rte-Engaging	-0.2565	0.0001	-0.2620	0.0001
fact_rte-Uses Knowledge	-0.3900	0.0001	-0.5263	0.0001
fact_rte-Overall	-0.2312	0.0001	-0.2360	0.0001
fact_stsb-Understandable	0.1994	0.0001	0.1999	0.0001
fact_stsb-Natural	0.1346	0.0106	0.1249	0.0178
fact_stsb-Maintains Context	0.1832	0.0005	0.1739	0.0009
fact_stsb-Engaging	0.4147	0.0001	0.4103	0.0001
fact_stsb-Uses Knowledge	0.4808	0.0001	0.4522	0.0001
fact_stsb-Overall	0.3324	0.0001	0.3220	0.0001

Table A.4: Correlation scores between the GLUE tasks on the conversation utterances evaluated against the knowledge base facts and the human annotator scores and their respective p-values on the **TopicalChat** dataset.

PersonaChat				
Predictor-Criteria	**Pearson's r**	**$p <$**	**Spearman's ρ**	**$p <$**
fact_mrpc-Understandable	0.1219	0.0349	0.1938	0.0007
fact_mrpc-Natural	0.0417	0.4721	0.0550	0.3425
fact_mrpc-Maintains Context	0.0252	0.6642	-0.0532	0.3589
fact_mrpc-Engaging	0.1302	0.0241	0.0461	0.4265
fact_mrpc-Uses Knowledge	-0.0046	0.9367	-0.0571	0.3247
fact_mrpc-Overall	0.0149	0.7972	-0.0726	0.2101
fact_qnli-Understandable	-0.1256	0.0296	-0.1494	0.0095
fact_qnli-Natural	-0.0642	0.2674	-0.0584	0.3131
fact_qnli-Maintains Context	-0.1478	0.0103	-0.1014	0.0794
fact_qnli-Engaging	-0.1157	0.0453	-0.0817	0.1583
fact_qnli-Uses Knowledge	-0.2733	0.0001	-0.2613	0.0001
fact_qnli-Overall	-0.1899	0.0009	-0.1734	0.0026
fact_qqp-Understandable	0.0476	0.4113	-0.0767	0.1850
fact_qqp-Natural	0.0762	0.1881	-0.0591	0.3072
fact_qqp-Maintains Context	0.0397	0.4936	0.0774	0.1813
fact_qqp-Engaging	0.1400	0.0152	0.1365	0.0180
fact_qqp-Uses Knowledge	0.2099	0.0003	0.4352	0.0001
fact_qqp-Overall	0.1613	0.0051	0.2230	0.0001
fact_rte-Understandable	-0.0296	0.6098	-0.0914	0.1142
fact_rte-Natural	0.0289	0.6181	0.0305	0.5993
fact_rte-Maintains Context	-0.0371	0.5225	-0.0041	0.9440
fact_rte-Engaging	-0.1091	0.0591	-0.0726	0.2100
fact_rte-Uses Knowledge	-0.4052	0.0001	-0.3481	0.0001
fact_rte-Overall	-0.1232	0.0330	-0.1122	0.0522
fact_stsb-Understandable	0.0250	0.6660	0.0307	0.5961
fact_stsb-Natural	0.0302	0.6025	-0.0032	0.9555
fact_stsb-Maintains Context	0.1537	0.0077	0.0876	0.1300
fact_stsb-Engaging	0.3419	0.0001	0.3378	0.0001
fact_stsb-Uses Knowledge	0.7329	0.0001	0.7173	0.0001
fact_stsb-Overall	0.3742	0.0001	0.3898	0.0001

Table A.5: Correlation scores between the GLUE tasks on the conversation utterances evaluated against the knowledge base facts and the human annotator scores and their respective p-values on the **PersonaChat** dataset.

TopicalChat				
Predictor-Criteria	**Pearson's r**	$p <$	**Spearman's ρ**	$p <$
lin-reg_pair-Understandable	0.1664	0.0015	0.1638	0.0018
lin-reg_pair-Natural	0.1986	0.0001	0.1972	0.0002
lin-reg_pair-Maintains Context	0.2859	0.0001	0.2946	0.0001
lin-reg_pair-Engaging	0.3884	0.0001	0.4008	0.0001
lin-reg_pair-Uses Knowledge	0.2455	0.0001	0.2751	0.0001
lin-reg_pair-Overall	0.3492	0.0001	0.3587	0.0001
lin-reg_fact-Understandable	0.2572	0.0001	0.2362	0.0001
lin-reg_fact-Natural	0.2244	0.0001	0.1805	0.0006
lin-reg_fact-Maintains Context	0.2272	0.0001	0.1921	0.0002
lin-reg_fact-Engaging	0.4358	0.0001	0.4078	0.0001
lin-reg_fact-Uses Knowledge	0.5517	0.0001	0.5182	0.0001
lin-reg_fact-Overall	0.3897	0.0001	0.3482	0.0001
lin-reg_single-Understandable	0.2542	0.0001	0.2470	0.0001
lin-reg_single-Natural	0.2148	0.0001	0.1853	0.0004
lin-reg_single-Maintains Context	0.0469	0.3753	0.0197	0.7094
lin-reg_single-Engaging	0.1483	0.0048	0.0881	0.0952
lin-reg_single-Uses Knowledge	0.0699	0.1855	0.0377	0.4754
lin-reg_single-Overall	0.1432	0.0065	0.1138	0.0308
lin-reg_all-Understandable	0.3420	0.0001	0.3390	0.0001
lin-reg_all-Natural	0.3357	0.0001	0.3130	0.0001
lin-reg_all-Maintains Context	0.3489	0.0001	0.3409	0.0001
lin-reg_all-Engaging	0.5335	0.0001	0.5364	0.0001
lin-reg_all-Uses Knowledge	0.5119	0.0001	0.5295	0.0001
lin-reg_all-Overall	0.4974	0.0001	0.4877	0.0001

Table A.6: Correlation scores between the combined GLUE scores with linear regression and the human annotator scores and their respective p-values on the **TopicalChat** dataset.

PersonaChat				
Predictor-Criteria	**Pearson's *r***	***p* <**	**Spearman's *ρ***	***p* <**
lin-reg_pair-Understandable	0.1626	0.0047	0.1345	0.0198
lin-reg_pair-Natural	0.2033	0.0004	0.2160	0.0002
lin-reg_pair-Maintains Context	0.4178	0.0001	0.3981	0.0001
lin-reg_pair-Engaging	0.2185	0.0001	0.2216	0.0001
lin-reg_pair-Uses Knowledge	0.1909	0.0009	0.1916	0.0009
lin-reg_pair-Overall	0.3975	0.0001	0.3802	0.0001
lin-reg_fact-Understandable	0.0255	0.6606	0.0328	0.5720
lin-reg_fact-Natural	0.0546	0.3456	0.0319	0.5823
lin-reg_fact-Maintains Context	0.1783	0.0019	0.1110	0.0548
lin-reg_fact-Engaging	0.3446	0.0001	0.3412	0.0001
lin-reg_fact-Uses Knowledge	0.6959	0.0001	0.7020	0.0001
lin-reg_fact-Overall	0.3990	0.0001	0.4135	0.0001
lin-reg_single-Understandable	0.0643	0.2668	0.0603	0.2978
lin-reg_single-Natural	-0.0285	0.6226	0.0302	0.6024
lin-reg_single-Maintains Context	0.0974	0.0922	0.1012	0.0801
lin-reg_single-Engaging	0.1675	0.0036	0.1597	0.0056
lin-reg_single-Uses Knowledge	0.0464	0.4230	0.0644	0.2661
lin-reg_single-Overall	0.1216	0.0353	0.1263	0.0287
lin-reg_all-Understandable	0.1214	0.0355	0.1218	0.0350
lin-reg_all-Natural	0.1728	0.0027	0.2044	0.0004
lin-reg_all-Maintains Context	0.4029	0.0001	0.3707	0.0001
lin-reg_all-Engaging	0.3272	0.0001	0.3306	0.0001
lin-reg_all-Uses Knowledge	0.5921	0.0001	0.5898	0.0001
lin-reg_all-Overall	0.5290	0.0001	0.5382	0.0001

Table A.7: Correlation scores between the combined GLUE scores with linear regression and the human annotator scores and their respective p-values on the **PersonaChat** dataset.

Bibliography

[1] V. Logacheva, M. Burtsev, V. Malykh, V. Polulyakh and A. Seliverstov, "ConvAI Dataset of Topic-Oriented Human-to-Chatbot Dialogues", *The NIPS'17 Competition: Building Intelligent Systems*, Springer, 2018 47 (cit. on pp. 1, 2, 35, 38, 44, 82).

[2] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, ed. by J. Burstein, C. Doran and T. Solorio, Association for Computational Linguistics, 2019 4171, URL: https://doi.org/10.18653/v1/n19-1423 (cit. on pp. 7, 17, 18, 30, 43, 45, 54, 55, 71, 80).

[3] A. Radford et al., *Language models are unsupervised multitask learners*, OpenAI blog **1** (2019) 9 (cit. on pp. 7, 17, 43, 45, 68, 72, 79–81).

[4] Z. Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding", *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, ed. by H. M. Wallach et al., 2019 5754, URL: http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding (cit. on pp. 7, 17, 43, 45, 46, 79, 80).

[5] A. Wang et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019, URL: https://openreview.net/forum?id=rJ4km2R5t7 (cit. on pp. 7, 31, 32, 54).

[6] R. Higashinaka et al., *Overview of dialogue breakdown detection challenge 3*, Proceedings of dialog system technology challenge **6** (2017) (cit. on pp. 8, 23).

[7] R. Higashinaka et al., "Overview of the Dialogue Breakdown Detection Challenge 4", *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, ed. by E. Marchi, S. M. Siniscalchi, S. Cumani, V. M. Salerno and H. Li, vol. 714, Lecture Notes in Electrical Engineering, Springer, 2019 403,

URL: https://doi.org/10.1007/978-981-15-9323-9%5C_38 (cit. on pp. 8, 23, 29, 67, 72–74).

[8] Y. Zhang et al., "DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, ed. by A. Çelikyilmaz and T. Wen, Association for Computational Linguistics, 2020 270, URL: https://www.aclweb.org/anthology/2020.acl-demos.30/ (cit. on pp. 8, 81).

[9] T. Szandala, *Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks*, CoRR **abs/2010.09458** (2020), arXiv: 2010.09458, URL: https://arxiv.org/abs/2010.09458 (cit. on p. 12).

[10] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning representations by back-propagating errors*, nature **323** (1986) 533 (cit. on p. 12).

[11] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems", *Advances in neural information processing systems*, 1997 473 (cit. on pp. 13, 17).

[12] K. Cho, B. van Merrienboer, D. Bahdanau and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches", *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, ed. by D. Wu, M. Carpuat, X. Carreras and E. M. Vecchi, Association for Computational Linguistics, 2014 103, URL: https://aclanthology.org/W14-4012/ (cit. on p. 13).

[13] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks", *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, 2014 3104, URL: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks (cit. on pp. 13, 16, 80).

[14] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, ed. by Y. Bengio and Y. LeCun, 2015, URL: http://arxiv.org/abs/1409.0473 (cit. on p. 14).

[15] A. Vaswani et al., "Attention is All you Need", *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, ed. by I. Guyon et al., 2017 5998, URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need (cit. on pp. 15, 17, 72, 80).

[16] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Studies in Natural Language Processing, Cambridge University Press, 2000 (cit. on pp. 15, 80).

[17] A. Gatt and E. Krahmer, *Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation*, J. Artif. Intell. Res. **61** (2018) 65, URL: https://doi.org/10.1613/jair.5477 (cit. on pp. 16, 80).

[18] S. Santhanam and S. Shaikh, *A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions*, CoRR **abs/1906.00500** (2019), arXiv: 1906.00500, URL: http://arxiv.org/abs/1906.00500 (cit. on pp. 16, 80).

[19] I. Sutskever, J. Martens and G. E. Hinton, “Generating Text with Recurrent Neural Networks”, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, ed. by L. Getoor and T. Scheffer, Omnipress, 2011 1017, URL: https://icml.cc/2011/papers/524%5C_icmlpaper.pdf (cit. on pp. 16, 80).

[20] O. Vinyals and Q. V. Le, *A Neural Conversational Model*, CoRR **abs/1506.05869** (2015), arXiv: 1506.05869, URL: http://arxiv.org/abs/1506.05869 (cit. on pp. 16, 36, 37, 80).

[21] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville and J. Pineau, “Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models”, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, ed. by D. Schuurmans and M. P. Wellman, AAAI Press, 2016 3776, URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957 (cit. on pp. 16, 24, 36, 37, 80).

[22] I. V. Serban et al., “A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues”, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, ed. by S. P. Singh and S. Markovitch, AAAI Press, 2017 3295, URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567 (cit. on pp. 16, 37, 80).

[23] Y. Park, J. Cho and G. Kim, “A Hierarchical Latent Structure for Variational Conversation Modeling”, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, ed. by M. A. Walker, H. Ji and A. Stent, Association for Computational Linguistics, 2018 1792, URL: https://doi.org/10.18653/v1/n18-1162 (cit. on pp. 16, 37, 80, 81).

[24] H. Chen, X. Liu, D. Yin and J. Tang, *A Survey on Dialogue Systems: Recent Advances and New Frontiers*, SIGKDD Explor. **19** (2017) 25, URL: https://doi.org/10.1145/3166054.3166058 (cit. on pp. 16, 35).

[25] F. Jelinek, *Continuous speech recognition by statistical methods*, Proceedings of the IEEE **64** (1976) 532 (cit. on p. 17).

[26] J. Baker, *The DRAGON system–An overview*, IEEE Transactions on Acoustics, speech, and signal Processing **23** (1975) 24 (cit. on p. 17).

[27] M. Peters et al., "Deep Contextualized Word Representations", *Proceedings of NAACL-HLP 2018*, 2018 2227 (cit. on pp. 17, 43, 45).

[28] G. Melis, C. Dyer and P. Blunsom, *On the state of the art of evaluation in neural language models*, arXiv preprint arXiv:1707.05589 (2017) (cit. on p. 17).

[29] S. Golovanov et al., "Large-scale transfer learning for natural language generation", *Proceedings of ACL 2019*, 2019 6053 (cit. on p. 17).

[30] Z. Dai et al., "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context", *Proceedings of ACL 2019*, Florence, Italy: Association for Computational Linguistics, 2019 2978, URL: https://www.aclweb.org/anthology/P19-1285 (cit. on p. 18).

[31] A. Conneau and G. Lample, "Cross-lingual Language Model Pretraining", *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, ed. by H. M. Wallach et al., 2019 7057, URL: https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html (cit. on pp. 18, 30).

[32] Y. Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, CoRR **abs/1907.11692** (2019), arXiv: 1907.11692, URL: http://arxiv.org/abs/1907.11692 (cit. on pp. 18, 26, 57, 73).

[33] C. Raffel et al., *Exploring the limits of transfer learning with a unified text-to-text transformer*, arXiv preprint arXiv:1910.10683 (2019) (cit. on p. 18).

[34] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong and R. Socher, *Ctrl: A conditional transformer language model for controllable generation*, arXiv preprint arXiv:1909.05858 (2019) (cit. on p. 18).

[35] I. Tenney et al., "What do you learn from context? Probing for sentence structure in contextualized word representations", *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019, URL: https://openreview.net/forum?id=SJzSgnRcKX (cit. on p. 18).

[36] X. Zhou, Y. Zhang, L. Cui and D. Huang, "Evaluating Commonsense in Pre-Trained Language Models", *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, 2020 9733, URL: https://ojs.aaai.org/index.php/AAAI/article/view/6523 (cit. on p. 18).

[37] V. Chandola, A. Banerjee and V. Kumar, *Anomaly detection: A survey*, ACM computing surveys (CSUR) **41** (2009) 15 (cit. on pp. 19, 36).

[38] R. Chalapathy and S. Chawla, *Deep learning for anomaly detection: A survey*, arXiv preprint arXiv:1901.03407 (2019) (cit. on pp. 19, 36).

[39] S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney and D. Song, *Anomalous Instance Detection in Deep Learning: A Survey*, CoRR **abs/2003.06979** (2020), arXiv: 2003.06979, URL: https://arxiv.org/abs/2003.06979 (cit. on pp. 19, 20, 68).

[40] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks", *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017, URL: https://openreview.net/forum?id=Hkg4TI9xl (cit. on pp. 19, 20, 49, 69, 70).

[41] S. Liang, Y. Li and R. Srikant, "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks", *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018, URL: https://openreview.net/forum?id=H1VGkIxRZ (cit. on pp. 19, 21, 69, 70).

[42] Y. Hsu, Y. Shen, H. Jin and Z. Kira, "Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data", *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, IEEE, 2020 10948, URL: https://doi.org/10.1109/CVPR42600.2020.01096 (cit. on pp. 19, 21, 69–71).

[43] J. Ren et al., "Likelihood Ratios for Out-of-Distribution Detection", *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, ed. by H. M. Wallach et al., 2019 14680, URL: http://papers.nips.cc/paper/9611-likelihood-ratios-for-out-of-distribution-detection (cit. on pp. 19, 21, 69, 71, 76).

[44] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 2004 74, URL: https://www.aclweb.org/anthology/W04-1013 (cit. on p. 23).

[45] K. Papineni, S. Roukos, T. Ward and W. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, ACL, 2002 311, URL: https://www.aclweb.org/anthology/P02-1040/ (cit. on p. 23).

[46] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, ed. by J. Goldstein, A. Lavie, C. Lin and C. R. Voss, Association for Computational Linguistics, 2005 65, URL: https://www.aclweb.org/anthology/W05-0909/ (cit. on p. 24).

[47] A. Ritter, C. Cherry and W. B. Dolan, “Data-driven response generation in social media”, *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2011 583 (cit. on p. 24).

[48] K. Yoshino et al., *Dialog System Technology Challenge 7*, CoRR **abs/1901.03461** (2019), arXiv: 1901.03461, URL: http://arxiv.org/abs/1901.03461 (cit. on pp. 24, 35).

[49] C. Liu et al., “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, ed. by J. Su, X. Carreras and K. Duh, The Association for Computational Linguistics, 2016 2122, URL: https://doi.org/10.18653/v1/d16-1230 (cit. on p. 24).

[50] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT”, *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020, URL: https://openreview.net/forum?id=SkeHuCVFDr (cit. on pp. 24, 25).

[51] R. Lowe et al., “Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses”, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, ed. by R. Barzilay and M. Kan, Association for Computational Linguistics, 2017 1116, URL: https://doi.org/10.18653/v1/P17-1103 (cit. on pp. 24, 25, 44).

[52] C. Tao, L. Mou, D. Zhao and R. Yan, “RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems”, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, ed. by S. A. McIlraith and K. Q. Weinberger, AAAI Press, 2018 722, URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16179 (cit. on pp. 25, 26, 44, 47).

[53] A. B. Sai, M. D. Gupta, M. M. Khapra and M. Srinivasan, “Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses”, *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii,*

USA, January 27 - February 1, 2019, AAAI Press, 2019 6220,
URL: https://doi.org/10.1609/aaai.v33i01.33016220 (cit. on p. 25).

[54] K. Sinha et al., "Learning an Unreferenced Metric for Online Dialogue Evaluation", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, ed. by D. Jurafsky, J. Chai, N. Schluter and J. R. Tetreault, Association for Computational Linguistics, 2020 2430,
URL: https://doi.org/10.18653/v1/2020.acl-main.220 (cit. on pp. 26, 27).

[55] A. B. Sai, A. K. Mohankumar, S. Arora and M. M. Khapra, *Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining*, Trans. Assoc. Comput. Linguistics **8** (2020) 810,
URL: https://transacl.org/ojs/index.php/tacl/article/view/2389 (cit. on pp. 26, 27, 44, 68).

[56] Y. Li et al., "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset", *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, ed. by G. Kondrak and T. Watanabe, Asian Federation of Natural Language Processing, 2017 986, URL: https://aclanthology.org/I17-1099/ (cit. on p. 26).

[57] S. Mehri and M. Eskénazi, "USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, ed. by D. Jurafsky, J. Chai, N. Schluter and J. R. Tetreault, Association for Computational Linguistics, 2020 681,
URL: https://doi.org/10.18653/v1/2020.acl-main.64 (cit. on pp. 26, 27, 44, 54, 55, 57–59, 63, 68).

[58] X. Gao, Y. Zhang, M. Galley, C. Brockett and B. Dolan, "Dialogue Response Ranking Training with Large-Scale Human Feedback Data", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, ed. by B. Webber, T. Cohn, Y. He and Y. Liu, Association for Computational Linguistics, 2020 386,
URL: https://doi.org/10.18653/v1/2020.emnlp-main.28 (cit. on pp. 27, 28).

[59] C. Zhang et al., "DynaEval: Unifying Turn and Dialogue Level Evaluation", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, ed. by C. Zong, F. Xia, W. Li and R. Navigli, Association for Computational Linguistics, 2021 5676, URL: https://doi.org/10.18653/v1/2021.acl-long.441 (cit. on pp. 28, 29).

[60] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, ed. by K. Inui, J. Jiang, V. Ng and X. Wan,

Association for Computational Linguistics, 2019 3980,
URL: https://doi.org/10.18653/v1/D19-1410 (cit. on p. 28).

[61] R. Higashinaka, K. Funakoshi, Y. Kobayashi and M. Inaba, "The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics", *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, ed. by N. Calzolari et al., European Language Resources Association (ELRA), 2016, URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/525.html (cit. on p. 29).

[62] H. Sugiyama, "Dialogue Breakdown Detection Using BERT with Traditional Dialogue Features", *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, ed. by E. Marchi, S. M. Siniscalchi, S. Cumani, V. M. Salerno and H. Li, vol. 714, Lecture Notes in Electrical Engineering, Springer, 2019 419, URL: https://doi.org/10.1007/978-981-15-9323-9%5C_39 (cit. on pp. 30, 67, 73).

[63] J. Shin, A. Dirafzoon and A. Anshu, "Context-enriched attentive memory network with global and local encoding for dialogue breakdown detection", *Proceedings of the Workshop on Chatbots and Conversational Agents*, 2019 (cit. on pp. 29, 67, 73).

[64] C. Wang, S. Kato and T. Sakai, "RSL19BD at DBDC4: Ensemble of Decision Tree-Based and LSTM-Based Models", *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, ed. by E. Marchi, S. M. Siniscalchi, S. Cumani, V. M. Salerno and H. Li, vol. 714, Lecture Notes in Electrical Engineering, Springer, 2019 429, URL: https://doi.org/10.1007/978-981-15-9323-9%5C_40 (cit. on pp. 30, 67, 73).

[65] Q. Lin, S. Kundu and H. T. Ng, "A Co-Attentive Cross-Lingual Neural Model for Dialogue Breakdown Detection", *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, ed. by D. Scott, N. Bel and C. Zong, International Committee on Computational Linguistics, 2020 4201, URL: https://doi.org/10.18653/v1/2020.coling-main.371 (cit. on pp. 30, 31, 67, 73, 74).

[66] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, ed. by D. Jurafsky, J. Chai, N. Schluter and J. R. Tetreault, Association for Computational Linguistics, 2020 8440, URL: https://doi.org/10.18653/v1/2020.acl-main.747 (cit. on pp. 30, 73).

[67] A. Warstadt, A. Singh and S. R. Bowman, *Neural network acceptability judgments*, arXiv preprint arXiv:1805.12471 (2018) (cit. on p. 31).

[68] R. Socher et al.,
"Recursive deep models for semantic compositionality over a sentiment treebank",
Proceedings of the 2013 conference on empirical methods in natural language processing,
2013 1631 (cit. on p. 31).

[69] A. Ghandeharioun et al.,
Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems,
arXiv preprint arXiv:1906.09308 (2019) (cit. on pp. 31, 54).

[70] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases",
Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005
(cit. on p. 32).

[71] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation",
Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017
1 (cit. on p. 32).

[72] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang,
"SQuAD: 100,000+ Questions for Machine Comprehension of Text",
Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,
2016 2383 (cit. on p. 32).

[73] I. Dagan, O. Glickman and B. Magnini,
"The PASCAL recognising textual entailment challenge",
Machine Learning Challenges Workshop, Springer, 2005 177 (cit. on p. 32).

[74] R. Bar-Haim et al., "The second pascal recognising textual entailment challenge",
Proceedings of the second PASCAL challenges workshop on recognising textual entailment,
vol. 6, Venice, 2006 6 (cit. on p. 32).

[75] D. Giampiccolo, B. Magnini, I. Dagan and B. Dolan,
"The third pascal recognizing textual entailment challenge",
Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing,
Association for Computational Linguistics, 2007 1 (cit. on p. 32).

[76] L. Bentivogli, P. Clark, I. Dagan and D. Giampiccolo,
"The Fifth PASCAL Recognizing Textual Entailment Challenge.", *TAC*, 2009 (cit. on p. 32).

[77] A. Williams, N. Nangia and S. Bowman,
"A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference",
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018
1112 (cit. on p. 33).

[78] H. Levesque, E. Davis and L. Morgenstern, "The winograd schema challenge", *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012
(cit. on p. 34).

[79] E. Dinan et al., *The Second Conversational Intelligence Challenge (ConvAI2)*,
CoRR **abs/1902.00098** (2019), arXiv: 1902.00098,
URL: http://arxiv.org/abs/1902.00098 (cit. on pp. 35, 38, 44, 72, 79, 82).

[80] S. Larson et al.,
"Outlier Detection for Improved Data Quality and Diversity in Dialog Systems",
Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers),
ed. by J. Burstein, C. Doran and T. Solorio, Association for Computational Linguistics, 2019 517, URL: https://doi.org/10.18653/v1/n19-1051 (cit. on pp. 36, 68).

[81] N. Hollenstein, N. Schneider and B. Webber, "Inconsistency detection in semantic annotation", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016 3986 (cit. on p. 36).

[82] D. Guthrie, L. Guthrie and Y. Wilks,
An unsupervised approach for the detection of outliers in corpora, Statistics (2008) 3409 (cit. on p. 36).

[83] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs",
Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, CMCL@ACL 2011, Portland, Oregon, USA, June 23, 2011, ed. by F. Keller and D. Reitter, Association for Computational Linguistics, 2011 76,
URL: https://www.aclweb.org/anthology/W11-0609/ (cit. on pp. 37, 81).

[84] M. Burtsev et al., "The first conversational intelligence challenge",
The NIPS'17 Competition: Building Intelligent Systems, Springer, 2018 25
(cit. on pp. 38, 44, 82).

[85] S. Zhang et al., "Personalizing Dialogue Agents: I have a dog, do you have pets too?",
Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers,
ed. by I. Gurevych and Y. Miyao, Association for Computational Linguistics, 2018 2204,
URL: https://www.aclweb.org/anthology/P18-1205/
(cit. on pp. 38, 44, 57, 82).

[86] K. Kann, S. Rothe and K. Filippova,
"Sentence-Level Fluency Evaluation: References Help, But Can Be Spared!",
Proceedings of the 22nd Conference on Computational Natural Language Learning, 2018 313
(cit. on pp. 44, 45).

[87] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing",
Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020,
ed. by Q. Liu and D. Schlangen, Association for Computational Linguistics, 2020 38,
URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6/
(cit. on pp. 44, 81).

[88] P. Lison and J. Tiedemann,
"OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles",
Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016, ed. by N. Calzolari et al.,

European Language Resources Association (ELRA), 2016, URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/947.html (cit. on p. 47).

[89] A. Wang, K. Cho and C. A. G. Scholar, *BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model*, NAACL HLT 2019 (2019) 30 (cit. on p. 50).

[90] S. Welleck, J. Weston, A. Szlam and K. Cho, "Dialogue Natural Language Inference", *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, ed. by A. Korhonen, D. R. Traum and L. Màrquez, Association for Computational Linguistics, 2019 3731, URL: https://doi.org/10.18653/v1/p19-1363 (cit. on pp. 54, 63, 88).

[91] T. Wolf et al., *HuggingFace's Transformers: State-of-the-art Natural Language Processing*, CoRR **abs/1910.03771** (2019), arXiv: 1910.03771, URL: http://arxiv.org/abs/1910.03771 (cit. on pp. 55, 68, 72).

[92] J. X. Morris et al., *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*, 2020, arXiv: 2005.05909 [cs.CL] (cit. on p. 55).

[93] K. Gopalakrishnan et al., "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations", *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, ed. by G. Kubin and Z. Kacic, ISCA, 2019 1891, URL: https://doi.org/10.21437/Interspeech.2019-3079 (cit. on p. 57).

[94] R. Lowe, N. Pow, I. Serban and J. Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems", *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, The Association for Computer Linguistics, 2015 285, URL: https://doi.org/10.18653/v1/w15-4640 (cit. on p. 57).

[95] X. Liu, P. He, W. Chen and J. Gao, "Multi-Task Deep Neural Networks for Natural Language Understanding", *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, ed. by A. Korhonen, D. R. Traum and L. Màrquez, Association for Computational Linguistics, 2019 4487, URL: https://doi.org/10.18653/v1/p19-1441 (cit. on pp. 63, 88).

[96] R. E. Banchs and H. Li, "IRIS: a Chat-oriented Dialogue System based on the Vector Space Model", *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, The Association for Computer Linguistics, 2012 37, URL: https://www.aclweb.org/anthology/P12-3007/ (cit. on p. 72).

[97] S. Lee et al., *Multi-domain task-completion dialog challenge*, Dialog system technology challenges **8** (2019) 9 (cit. on p. 72).

[98] L. F. D'Haro et al., *Overview of the seventh Dialog System Technology Challenge: DSTC7*, Comput. Speech Lang. **62** (2020) 101068, URL: https://doi.org/10.1016/j.csl.2020.101068 (cit. on p. 79).

List of Figures

List of Tables