Shhh...!

Silence in dialogue

Lars Štěpán Laichter

Master's Thesis

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

DEPARTMENT
OF LOGIC
Faculty of Arts
Charles University

Title: Silence in dialogue

Author: Štěpán Lars Laichter

Birth date: 8. 12. 1994

Birth place: Prague, Czech Republic

Defense date: September 2022

Supervisor: Dr. Raquel Fernández

Degree: MA in Logic

The investigations were supported by
the Specific University Research (SVV) grant
from the Charles University in Prague.

Cover design generated by DALL·E
using the prompt `painting of a robot facing a wall in the style of Vilhelm Hammershøi`.

Contents

Abstract vii

Acknowledgments ix

1 Introduction 1

2 From silence to dialogue 5

2.1 A brief history of dialogue 5

2.1.1 Defining dialogue 5

2.1.2 Antiquity 6

2.1.3 Modernity 7

2.2 Computerization of dialogue 8

2.2.1 Rule-based dialogue systems 8

2.2.2 Statistical data-driven machine learning dialogue systems 11

2.2.3 Neural end-to-end learning dialogue systems 12

2.3 The end of history 13

3 Theories of dialogue 15

3.1 Formal pragmatics 16

3.1.1 Speech act theory 17

3.1.2 Conversational maxims 17

3.2 Conversation analysis 19

3.3 Talk-in-interaction theory 21
3.4 Dialogue as joint action 21
3.5 The interactive alignment account 23
3.6 The interactive stance 25
3.7 Conclusion 26

4 Learning silence **27**
4.1 Problem formulation 28
4.2 Model 29
4.3 Experimental set-up 30
4.3.1 Data 30
4.3.2 Fine-tuning 34
4.3.3 Evaluation 34
4.4 Experiments 36
4.4.1 Movie-level experiment 36
4.4.2 Director-level experiment 39
4.4.3 Multi-director-level experiment 43
4.4.4 Impact of fine-tuning dataset size 45
4.5 Discussion 46
4.5.1 Movie-level result 46
4.5.2 Director-level results 47
4.5.3 Multi-director-level results 47
4.5.4 Remaining questions 48
4.6 Future work 49
4.6.1 Human evaluation 50
4.6.2 Incremental speech generation 52
4.7 Overview 53

5 Conclusion **55**

5.1 Limitations 56

5.2 Ethics statement 57

Bibliography **59**

Abstract

Silence is an indispensable aspect of dialogue. The following thesis examines the silence in dialogue from a variety of perspectives. First, I provide a background on the historical development of theories of dialogue and the place of silence within them. Second, I conduct a study of the capacity of one of the most prominent contemporary language models, called the GPT-3, to model silence in dialogue. I fine-tune the model on a dataset based on movie subtitle data. I evaluate its performance on its capacity to infer the length of silence between subtitle pairs. The experiment proposes a method of fine-tuning the language model via silence encoded as character strings. The results show that GPT-3 fine-tuning can indeed improve the model's performance by inferring silence gaps between subtitle turns.

Keywords: dialogue, silence, GPT-3, fine-tuning, language models

I declare that I have written my diploma thesis independently and that I have properly cited all the sources and literature used, and that the work has not been used in the context of another university study or to obtain another or the same degree.

Abstrakt

Ticho je nezbytnou součástí dialogu. Následující diplomová práce zkoumá ticho v dialogu z teoretické a aplikované perspektivy. Nejprve mapuje historický vývoj teorií dialogu a jejich přístup k tichu. Dále zahrnuje studii schopnosti jednoho z nejvýznamnějších současných jazykových modelů GPT-3 modelovat ticho v dialogu. Model je laděn na datasetu založeném na datech titulků z filmů. Jeho výkon je hodnocen na základě schopnosti odhadovat délku ticha mezi dvojicemi titulků. Experiment navrhuje metodu ladění jazykového modelu pomocí ticha zakódovaného jako řetězce znaků. Výsledky ukazují, že ladění GPT-3 skutečně může zlepšit výkon modelu při odhadování mezer ticha mezi řadami titulků.

Keywords: dialog, ticho, GPT-3, ladění, jazykové modely

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Acknowledgments

I am very grateful to Prof. Dr. Raquel Fernández for her help in writing of this thesis. I also would like to thank Alexandre Kabbach for his comments and suggestions. I would also like to thank Valerie Fowles for proofreading and writing feedback.

Amsterdam Lars Š. Laichter

June, 2022.

CUNI Compulsory Statement:

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Chapter 1

Introduction

Hello there,

...

...

...

...

...

...

...

...

...

...

...

...

if this was a conversation, I would have just stayed silent for far too long. The following thesis is an exploration of the role of silence in dialogue, as well as of what it takes to teach a computational system to reproduce silence in dialogue in a human-like way.

I am writing this work against the backdrop of rapid development in the field of

large language models when computational systems outperform previous expectations of what was possible almost every couple of weeks. My interest in this field is fueled by curiosity and fascination with what is possible to achieve with conversational systems, as well as by the frustration and dissapointment by the lack of appreciation for the complexity and depth of something so fundamental as human dialogue. While there are important ethical concerns with the use of large language models (further discussed in section 5.2), there are also potentially huge societal gains to be made by achieving human-level computer-generated dialogue (Floridi and Chiriatti, 2020). These might range from making interacting with computers more accessible to people with disabilities, to enabling the development of dialogue-driven artificially intelligent personalised educational assistants.

The main takeaway of this thesis is that while current language models are a powerful technology, more work is necessary to establish the theoretical foundations to understanding dialogue, as well as to develop the training datasets and evaluation methods necessary for the technology to be used to its highest potential in producing good human-like dialogue.

I focus in my exploration on silence, as there is no question that it is an indispensable aspect of dialogue. While silence might appear to be an uninteresting void constituting the less important part of how humans talk to another, its use varies wildly among cultures, contexts, and individuals (Duhoe and Giddi, 2020; Lestary et al., 2018a; Stivers et al., 2009). In some cases, silence might be a simple feature of taking turns, and in other instances, it might be purposefully used to represent a particular state and sentiment of the speaker (Ibrahim and Muhammad, 2021).

Despite the importance of silence in dialogue, it is often neglected and treated as secondary in the implementation of conversational systems. Understanding the use of silence in dialogue is not only informative, as to help us better understand conversational exchanges between humans, but it can also inform the design and development of computational conversational systems. If silence is then used in

computational conversational systems properly, it can help to make these systems more useful and pleasant to interact with. As far as the aspiration for a full-fledged human-to-machine dialogue, it appears difficult to achieve without a sophisticated use of silence.

My thesis maps out and adds to the current understanding of silence in dialogue. The text is constituted by two main goals: (1) to map out various theories of dialogue and their treatment of silence, and (2) to conduct an experiment on the capacity of a state-of-the-art language model to model silence in dialogue. To do so, I propose an approach utilising subtitle datasets as a means of generating training datasets for GPT-3 fine-tuning.

The main contributions of the thesis include, firstly, an overview of some of the most prominent theories of dialogue. As there is not really a field focused on the study of dialogue, these theories are scattered across different disciplines and not frequently presented as in conversation with one another (Anderson et al., 2003; Ephratt, 2011). This makes this thesis an accessible entry point for those interested in orienting themselves in the different approaches to studying dialogue. Secondly, I present a study of the modelling of silence in dialogue as a case study of an aspect of dialogue that is not often paid attention to in the development of conversational systems. I propose a novel framework for the training of silence in dialogue via movie subtitles. This case study allows me to showcase the limitation of current large language models, as well as show the need for more sophisticated training datasets and evaluation methods.

The broad goal of the thesis is to outline and provide a foundation for anyone interested in implementing conversational systems featuring more sophisticated uses of silence, as well as anyone interested in gaining a better understanding of the use of silence in dialogue in general.

Chapter 2

From silence to dialogue

2.1 A brief history of dialogue

The way in which humans engage in dialogue has been subject to multiple notable transformations. The transition from spoken to written dialogue, the transformation from face-to-face to remotely facilitated, and from human-only to computer-mediated dialogue. The latter is the most recent and perhaps the most complex. The way in which we engage in dialogue has evolved with each of these transformations and with it also the understanding of what dialogue is about. As the understaing of dialogue grew, the study of dialogue progressed from a literary device to a subject of theoretical examination in computer science, linguistics, and other related disciplines. In the following exposition, I provide a brief overview of the transformations it has undergone. I pay particular attention to the formation of the first theories of dialogue, which can help elucidate the role of silence in dialogue.

2.1.1 Defining dialogue

What is a dialogue? Dialogue is commonly defined as a conversation between two or more persons or a similar exchange between a person and something else, such

as a computer (Merriam-Webster, 2022). One may notice that this definition is acutely circular, given that a conversation in this context is synonymous with dialogue. Nevertheless, the definition makes clear that dialogue is most frequently some form of exchange, be it spoken, written, or in some other form. Exchange in this context relies on an implied notion of personhood, assuming that dialogue requires two or more interlocutors. Arguably, the above-mentioned definition is far from capturing the essence of such a complex activity. The following exploration of silence in dialogue is a dive into one of the more opaque aspects that make up dialogue as such.

2.1.2 Antiquity

The history of the study of dialogue dates at least as old as ancient Greece. The word 'dialogue' itself etymologically comes from the Greek word *dialogos*. *Logos* is translated as 'word', or 'the meaning of the word'. And *dia* means 'through'. Thus, one way of interpreting the origin of the words is 'the stream of meaning' (Bohm et al., 2004).

Ancient Greece is a particularly interesting example, as dialogue was not only a literary form, but was a foundation to the formation of democracy (Goldhill et al., 2008). Perhaps the most extensive instantiation of this, among other literary works of that period, are Plato's dialogues.

Plato is particularly relevant, as his works are based on a consistent notion and treatment of dialogue. His dialogues are characterised by discourse composed of questions and answers on a philosophical or political topic. This has come to be known as the Elenchus or the Socratic method. Elenchus is designed to bring out assumptions about a particular topic between two interlocutors to arrive at a contradiction (Gregory, 1983). Given the consistent use of dialogue in Plato's work, the Elenchus can be considered in a way the first formulation of a theory of dialogue. It has also subsequently led to the developments of formal methods,

such as dialogical logic (Clerbout and McConaughey, 2022).

While dialogue in literature almost disappeared in the Christian empire in late antiquity, it has resurfaced in modern times as a popular literary device (Goldhill et al., 2008). Nonetheless, it is hard to point out authors who have used dialogue with the methodological rigour and wide cultural impact of the ancient Greek scholars.

2.1.3 Modernity

Modern times have witnessed an explosion in the variety of dialogue. This can mainly be attributed to the ubiquity of the printing press, radio, cinema, television, and later the internet. Dialogue has become an indispensable part of the toolkit of most writers.

There have been also some notable philosophical works which have analysed the use of dialogue. This includes the work of Paulo Freire (Goodson and Gill, 2014), who championed the use of dialogue as an education tool, and the work of the physicist David Bohm, who designed a framework for the use of dialogue to address societal problems (Bohm et al., 2004). There is also the work of the philosopher Mikhail Bakhtin, who significantly contributed to the understanding of dialogue within the literary context (Bakhtin, 2010). While the methodologies of all these authors are quite distinct, their works are considered attempts at holistic treatments of dialogue.

An important aspect in the development of understanding dialogue in the present times has been the computerization of dialogue. Given the nuance and importance of this development, I dedicate a whole section to this topic.

2.2 Computerization of dialogue

A major turning point in the history of dialogue was the emergence of the idea of engaging in a dialogue with a machine. While the idea of talking to a machine has origins preceding computers, turning the idea into reality only became fully feasible with the invention of universal computation in the mid-20th century. Since then, the field of dialogue systems has seen steady progress facilitated by different approaches that have produced varying levels of success. McTear (2020) groups the development into their phases: (1) Rule-based dialogue systems, (2) Statistical data-driven machine learning dialogue systems, and (3) Neural end-to-end learning dialogue systems.

The development of the computational approaches has been mostly driven by the intent to produce credible emulation of dialogue by a computational system. Therefore, the quality of dialogue itself often remained secondary and the goal was not to understand dialogue as a phonomenon of human behaviour in itself. Nevertheless, understanding of the scope of these technologies is important in assessing the possibilities of computational dialogue systems.

2.2.1 Rule-based dialogue systems

Rule-based dialogue systems feature a set of rules that determine the system's behaviour. This approach was characteristic of some of the earliest dialogue systems and remains in use, primarily in commercial applications.

McTear (2020) lists various limitations of the rule-based approach to dialogue systems. Some of these include (1) the propensity of these systems to fail due to deviations from the predefined conversation path, (2) the difficulty of scaling to domains that the system is not designed for, and (3) the difficulty of guaranteeing that the system is optimal. Despite these shortcomings, the rule-based system still enjoys wide popularity and some of the systems in this area can feature quite extensive complexity.

The Turing test

The idea of a rule-based dialogue system dates back at least to the Turing test, proposed by Alan Turing in the 1950s (Turing, 1950). The Turing test originally titled the "The Imitation Game", was designed to serve as a means to determine whether "machines can think". In its essence, the test asks a human to determine whether they are conversing with a human or a computer. In the original test, the conversation is conducted via messages written on a tape. Although Turing thought that computers would soon do as well as humans on this test, it has proven more challenging for computers to match human performance than anticipated (Moor, 2001). While the Turing test remains a seminal benchmark for computer generate conversation, the game proposed by Turing did not make any particular assumption on the design or machine implementation of the dialogue itself.

ELIZA

The first full-fledged rule-based dialogue system, ELIZA, was developed in a lab at MIT led by Joseph Weizenbaum in the 1960s (Weizenbaum, 1966). The particular success of ELIZA was its capacity to demonstrate how natural it is for humans to start treating a computer program as an acceptable counterpart in a dialogue. The interaction was driven by pre-pared scripts modified by a set of rules. While there have been more sophisticated dialogue systems following ELIZA, it remains an example of one of the first attempts of computerised dialogue. ELIZA exemplifies a simple rule-based dialogue strategy attempting to reproduce credible dialogue fragments to uphold the overall impression of being human-like.

Pask's conversation theory

Following the success of ELIZA, there were some attempts within the so-called cyberneticists community of the 1970s to formulate a theory of conversations for the development of dialogue systems. One particular theory was formulated by Gordon Pask (Pangaro, 2017; Pask, 1976). Because this theory was primarily developed with the goal of advancing computer dialogue systems and not to develop a theory of dialogue as a part of human behaviour, it falls primarily under approaches to machine dialogue and not the theories of dialogue to be discussed later.

Pask proposes his criteria for conversation which are primarily concerned with change. For a conversation to occur, there must be a change in one of the cognitive agents engaged in the conversation, such as a change in understandings, concepts, intent, and values (Pangaro, 2017). If there is no change present, it is just an exchange of messages. Pask's primary interest was to use this theory to develop educational software that could aid learning (Pask, 1975). Some of the ideas he even implemented. Nonetheless, his approach was utimately hindered by limitations common to other rule-based approaches to the development of dialogue systems.

Commercial applications

As mentioned earlier, the use of rule-based dialogue systems has enjoyed a particular level of popularity in commercial applications. This is primarily due to the reliability and predictability of these systems. The approach, however, has been augmented with specific terminology and additional approaches, namely the use of intents and entities (Lorenc, 2021). This allows the rule-based systems to adopt some of the features of more statistical data-driven systems, which are discussed in the following section while retaining some of the advantages of the rule-based approach. Entities and intents are particularly useful in goal-driven

dialogues, as intent generally captures the goal of the interlocutor and entity of the object that is being talked about (Nadeau and Sekine, 2007; Junmei and William, 2019). Intent classification and entity extraction have become important technical problems on their own.

2.2.2 Statistical data-driven machine learning dialogue systems

Statistical data-driven machine learning dialogue systems have been developed to overcome the limitations of rule-based systems. These dialogue systems are trained on large corpora of data, based on the belief that dialogue can be produced by replicating the patterns that exist in the language in these datasets (McTear, 2020).

The statistical aspect of this approach means that the various components of the systems are modelled probabilistically (McTear, 2020, p. 72). This allows for the systems to better manage uncertainty and to be more robust to changes in the conversation. The data-driven driven aspect of this approach means that the statistical models are trained on various datasets. This data can come from various sources, such as past conversations, language corpora, or data from interactions between real and simulated users.

Core principles

One characteristic feature of this approach is extensive use of grammar parsers to be able to better understand the received input (Taylor et al., 2003). Based on the input, the system can produce an output based on the statistical patterns that exist in the dataset it is trained on. For example, in the case of corpus-based dialogue management, the system tries to find the most likely response to the user's input based on the preceding turns of dialogue.

The challenge for this approach is that the number of preceding states of the

dialogue can be very high, or the corpus might not be large enough to provide the needed replies to all possible turns in a dialogue (McTear, 2020, p. 77).

While there exists a series of other strategies for dialogue management (example-based models (Lee et al., 2009), Hidden Markov models (Cuayáhuitl et al., 2005), Bayesian networks (Meng et al., 2003) etc.), the reinforcement learning approach is perhaps the most notable one.

Reinforcement learning

Reinforcement learning is based on the notion of a state environment which is being explored by the agent. Each state is associated with a reward. The agent must choose between a range of options and choose the one that maximizes the reward until it reaches a final state. The goal is to find an optimal policy that maximises the expected reward (McTear, 2020, p. 81). The use of this principle in dialogue is such that dialogue is treated as a reward space and each turn of the conversation is associated with a reward (Scheffler and Young, 2002; Frampton and Lemon, 2005).

While reinforcement learning in dialogue produced some promising results, the challenge for these systems is an increasing number of user goals. This leads to a very large space of possible dialogue states, making exact dialogue state updating untractable (McTear, 2020, p. 87).

2.2.3 Neural end-to-end learning dialogue systems

Neural end-to-end learning dialogue systems leverage large amounts of data to produce models which do not require fine-tuning of individual components. In comparison to previously described approaches, an input utterance is mapped directly to an output response without requiring any processing by the modules of the traditional modularised architecture (McTear, 2020, p. 125). This mapping is generally done without an intermediate explicit representation, also referred to

as sequence-to-sequence mapping (Seq2Seq) (Sutskever et al., 2014).

The processing and representing of input is known as encoding while generating output is known as decoding. In most cases, the encoding is done via word embedding which converts the linguistic input into a unique real-number vector. The representation of the model in terms of one unified vector space constitutes a major advantage of the approach, as it allows for a fine-tuning of the system as a whole rather than the individual components. Due to this advantage, this approach has vastly outperformed previously existing dialogue systems.

While this approach has produced some impressive outputs, it remains limited since neural dialogue systems require vast amounts of data for training. Moreover, the need for large datasets has mostly limited the research to large companies and labs that can afford the costs of acquiring this data and computational resources to train the models. Examples of some of these models include Google's Meena, Facebook's BlenderBot, and OpenAI's GPT-3 (Komeili et al., 2022; Brown et al., 2020).

2.3 The end of history

The development of understanding dialogue has culminated in an ever-increasing capacity to mechanistically reproduce dialogue. This has come about mainly through experimentation rather than by developing an overall theory of dialogue. None of the approaches to constructing or modelling dialogue duly considered silence in dialogue. In the literary context, silence is omitted, except for generalised instances, such as the use of ellipsis. When it comes to the computerization of dialogue, most computational systems have not prioritised silence. In the next chapter, I map out various theories of dialogue and discuss their possible explanative power regarding silence in dialogue.

Chapter 3

Theories of dialogue

The majority of machine-generated dialogue currently falls short of its human-produced counterpart. Although there is an ever-increasing number of humans engaging with machine-generated dialogue through virtual assistants, chatbots, and other dialogue interfaces, the development of such technologies is rarely grounded in theories of dialogue. This holds for the implementation and accounts of silence in dialogue as well. In the following chapter, I explore theories which could potentially inform and ground the development of computational dialogue systems, as well as the subsequent question of understanding silence in dialogue.

Despite the importance of dialogue, both human-to-human and machine-to-human dialogue in our daily lives, the landscape of dialogue theory remains rather sparse, scattered with various attempts to describe the phenomenon of dialogue across different academic disciplines, ranging from logic to psycholinguistics. In this chapter, I (1) survey the currently available theories and (2) discuss their relevance to silence in dialogue.

What qualifies as a theory of dialogue?

For the following review, a theory of dialogue is a body of arguments that aim to answer questions, such as *What is a dialogue? What is its purpose? How should*

it be studied? How can it be emulated computationally? A theory of dialogue should be, among other things, grounded in empirical evidence and be coherent with the naive understanding of its daily use.

The core motivation for the search for a theory of dialogue can be well summed up by Kant's quote "practice without theory is blind" (Murphy, 1998). This is to say that while we are engaging in dialogue all the time, without a proper theory, we are not able to fully understand its principles. In many cases, we lack the appropriate conceptual apparatus to describe various properties of dialogue in general terms. Moreover, as the prevalence of human-to-machine dialogue increases, we must think about how to design the dialogues that so many people end up engaging with. Without the attempt to formulate a theory of what constitutes a dialogue and by which principles it should be constructed, we are only fumbling in the dark.

This further applies to an understanding of silence in dialogue. A theory which does not account for the role of silence in dialogue cannot provide a complete account of what dialogue is.

3.1 Formal pragmatics

Formal pragmatics was among the first fields to lay a foundation for a systematic study of dialogue. The goal of formal pragmatics is to understand how context determines the meaning of utterances. Context is then viewed as that which changes from utterance to utterance. One could say that the study of formal pragmatics focuses on the overlap between semantics and pragmatics.

The field has been pioneered in the 1950s by the work of J. L. Austin, John R. Searl and Paul Grice (Grice, 1989a; Austin and Warnock, 1962). It is predicated on the general observation that interpretation of utterances remains remarkably consistent between speakers (Potts, 2009). Although formal pragmatics originates in philosophy, its findings have drawn upon and impacted many other fields.

The formal pragmatists made an important observation that focusing only on utterances makes them mostly incomplete, as they largely depend on and interact with context. The context which influences the meaning of utterances includes the silence which precedes and follows an utterance. This makes formal pragmatics a relevant discipline to review before engaging with the topic of silence in dialogue.

3.1.1 Speech act theory

Inspired by the work of Austin (1975), Searle developed a taxonomy of illocutionary acts (Searle, 1975). Generally speaking, these are speech acts which describe what was done. Searle proposed to classify them within five distinct classes: (1) Representative or assertive, (2) Directive, (3) Commissive, (4) Expressive, and (5) Declarative. While the nuances of the individual classes do not bear that much importance and do not apply to dialogue directly, the attempt at the classification of various speech acts is an example characteristic of the speech act theory and early attempts at utterance classification.

The understanding of Searle that various aspects of speech can be classified is a precursor to attempts to classify the use of silence in dialogue. For example, Bruneau (1973) classifies silence as (1) psycholinguistic silence, (2) interactive silence, (3) sociocultural silence. While the specifics of these classes are not particularly relevant, they exemplify one prominent approach to understanding an aspect of dialogue.

3.1.2 Conversational maxims

Following the attempt by Searle to classify speech acts, Grice (1989b) provides a set of maxims providing principles for rational conversation. Grice motivates these maxims with the notion of conversational implicature, which is meant to allow reasoners to construct an inferential connection between what is meant and

what is implied. The maxims include:

1. **Quantity:** The quality of a conversation is determined by the utterances being neither more nor less of what is required.

2. **Quality:** The quality of a conversation is determined by the utterances being genuine and not spurious.

3. **Relation:** The conversational contributions should be appropriate to the immediate needs at each stage of the exchange.

4. **Manner:** It is clear what contributions to the conversation are made and the contributions are delivered appropriately.

The limitation of Grice's maxims is that they apply to only a narrow case when it comes to dialogue. Rational dialogue is such in which the participants subscribe to the goals that Grice assumes for dialogue, such as giving and receiving information, influencing and being influenced by others (Grice, 1989b, p. 30). Although the paper is not particularly formal, it marks the inception of the more formal approaches to pragmatics.

What is interesting about Grice's conversational maxims is that they are normative—implying that a "good dialogue" should meet particular characteristics. While such an approach can appear imposing, dialogue can frequently be carried out based on implied maxims. When it comes to silence, these maxims can be highly culturally dependent. For example, various cultures are claimed to have different tolerance for the length of silence (Duhoe and Giddi, 2020). Thus, given this implied maxim, for a "good dialogue" the interlocutors should follow their cultural norms when it comes to the permissible length of silence. The study of maxims in dialogue is, therefore, another available strategy for the study of dialogue.

3.2 Conversation analysis

Conversation analysis has emerged from sociolinguistics as a means of analysisng of common human interactions. There are various contributions of conversation analysis which have had a lasting impact on the discipline and beyond when it comes to understanding dialogue. In particular, it is the approach to structuring conversation which can be divided into three parts. Firstly, it concerns when a speaker decides to speak during a dialogue (i.e. turn-taking, dialogue repair, etc.). Secondly, it concerns how the utterances of an individual speaker relate to one another (i.e. adjacency pairs, etc.). Thirdly, it concerns the different functions of dialogue, such as establishing roles, etc. While the taxonomy provided by conversation analysis is quite extensive, I am including at least some of the most prominent examples to illustrate the diversity of the different parts of the approach.

1. **Turn-taking**: Turn-taking refers to the phenomenon of interchangeably assuming the role of a speaker or listener in conversation (Levinson, 1983). It does not refer only to the sequence of the interchanging of roles, but it also analyses the overlap and gaps that naturally occur. (Sacks et al., 1978) suggest that turn-taking is governed by quite an elaborate set of rules which can be further analysed. In their view, turn-allocational techniques are distributed into two groups: (1) those in which the next turn is allocated by the current speaker selects a next speaker and (2) those in which the next turn is allocated by self-selection.

2. **Sequence organisation**: Sequence organisation is the insight that dialogue is organised in a particular order which follows the order of related communicative actions. This is mainly done through sequential organisation constituted of adjacency pairs (Levinson, 1983). Adjacency pairs can be thought of as a basic building block of dialogue. Each adjacency pair consists of two

utterances produced by a different speaker. Examples include questions-answers, offer-acceptance and refusal, and compliment-response (Sacks et al., 1978).

3. **Repair**: Repair is a process introduced in conversation as a way for a speaker to identify and correct an error in what has been uttered and restate the utterance with some sort of a correction (Levinson, 1983). The term was first published by Fromkin (1971). In the spoken context, this might include an instance in which a speaker fails to make himself audible or comprehensible to a recipient. There are multiple combinations of self-initiated or other-initiated repair.

Overall, conversation analysis provides a detailed account of situated dialogue supported by a long-standing tradition of empirical research. However, the focus on a detailed account of conversation in different situations also seems to be its shortcoming. The approach overvalues individual instances of dialogue at the expense of formulating a general theory. In addition, it is be overly focused on face-to-face spoken dialogue while not paying as much attention to other forms of dialogue, such as human-to-machine dialogue. It is hard to say how well would the analysis, given its emphasis on detail, generalises to other types of dialogue.

Conversation analysis is, nevertheless, particularly relevant to the study of silence. The study of turn-taking lends itself well to examining the role of silence in between turns. There is a rich scholarship on turn-taking analysis, which sometimes also focuses on the role of silence. For example, Lestary et al. (2018b) investigates the purposes behind interruptions and the meaning of silence in conversation from the perspective of conversation analysis. Their methodology consists of analysing conversation transcripts to identify reasons for interruptions and the impact of silence on conversation flows.

3.3 Talk-in-interaction theory

In his book, *Approaching Dialogue*, Linell draws on empirical results from a wide range of fields to develop and support his theory of talk-in-interaction (Linell, 1998). These fields include discourse analysis, interactional linguistics, conversation analysis, ethnomethodological analysis of talk, symbolic interactionism, and communication science. However, conversation analysis remains the most dominant in his approach. His analysis of the talk-in-interaction starts with consideration of the most local features of dialogue and spans to macro features of communication.

Linell defends a theory called dialogism as a tenable account of cognition, individually-based information processing, communication, and language as a code. The dialogical theory views the brain as a sense-making system and it emphasizes the role of the other, as well as interactions and contexts. The theory challenges notions of individualism and rather assumes that sense-making occurs in communication and interventions with the world. The theories are primarily grounded in different accounts of continental philosophy. Although the whole book is often too concerned with individual examples rather than developing the overall theory, it provides a detailed account of the historical development of dialogism.

Linell does not hold particular views on silence. However, one can infer his possible treatment of silence based on his view of langue which he defines as "a stock of linguistic resources, i.e. expressions with associated semantic representations (abstract or decontextualized meanings) which are integrated within systems" (Linell, 1998, p. 3-4). These resources include silence. Therefore, Linell would likely view silence as one of the "codes" that one can manipulate through formal mechanisms.

3.4 Dialogue as joint action

Clark (1996) approaches the development of a theory of dialogue from the perspe-

ctive of a psycholinguist. In his book, *Using Language* Clark develops a theory that dialogue is a form of joint action. He takes joint action to be one carried out by an ensemble of people acting in coordination with each other. He distinguishes between personal and nonpersonal settings to make a difference between dialogue and monologue. He also introduces other types of settings, which include personal, nonpersonal, institutional, prescriptive, fictional, mediated, and private (Clark, 1996, p. 8).

In his view, langue is more than speakers speaking and listeners listening. It is the joint action that emerges when speakers and listeners, writers and readers perform their actions in coordination, as ensembles. There are several core propositions that he makes about language throughout to book: (1) language fundamentally is used for social purposes, (2) language use is a species of joint action, and (3) language use always involves the speaker's meaning and the addressee's understanding, (4) the basic setting for language use is face-to-face conversation, (5) language use often has more than one layer of activity, (6) the study of language use is both a cognitive and social science.

His approach is contrary to the popular notion within cognitive science, where language is seen as an individual and social process within social sciences. He emphasises that space and social setting play an important role, as a form of joint activity. He sees joint activity to be conducted mostly through joint actions. In other words, Clark argues that language use embodies both individual and social processes. Furthermore, he emphasises the importance of common ground, as people cannot take joint actions without assuming some common ground (Clark, 1996, p. 120).

Clark's theory features the breadth needed for contemporary dialogue use while remaining grounded within other scientific fields, namely social science and cognitive science. It seems well-positioned to have sufficient explanatory power for individual instances in dialogue, while also painting a picture of the overall dynamics of dialogue. My main objection is that dialogue does not always have

to be a matter of coordination, as Clark proposes. There are instances of dialogue that inhibit coordination between people. Furthermore, dialogue is for Clark only a subset of his theory could potentially.

When it comes to silence, Clark speaks of hiatus in fluent speech which he takes to be filled by more than just silence. He lists six common types of content that can be found in between turns: (1) no pause, (2) pause, (3) filler, (4) editing expression, (5) elongation, and (6) iconic gesture (Clark, 1996, p. 262). He takes this content to function as a signal that aids the coordination of speakers. Furthermore, he discusses silence as an issue that might hinder coordination. This might be the case when the silence grows too large (Clark, 1996, p. 269). Other than that, Clark does not attribute any particular meaning to silence in his theory.

3.5 The interactive alignment account

Pickering and Garrod (2004) propose the interactive alignment model of dialogue. Their account argues that linguistic representations employed by interlocutors in conversation align on various levels due to a set of mostly automatic processes. These automatic processes include mechanisms, such as priming, routines, and simple inference mechanisms. The analysis falls primarily under the field of psycholinguistics.

The premise of the paper is that often, theories of language are based on monologue. However, they claim dialogue to be the basic skill that can be engaged by children and illiterate people, while monologue requires some additional conceptual toolkit. Thus, they claim that dialogue is better suited to provide a mechanistic theory of langue. The challenge with dialogue is that it is inherently interactive and contextualised, which is perhaps the reason why it has been treated as secondary in the psycholinguistic discourse. Nevertheless, the proposal also applies to monologue, whici it treats as a special case of dialogue.

The interactive alignment model builds on analysis by Clark (1996), but the

authors also distance themselves from his analysis by claiming that he barely touches upon the processes behind generation and comprehension which constitute their main focus. While Clark takes coordination to be a moment when interlocutors conduct a joint activity, Pickering and Garrod take coordination as occurring when the interlocutors share the same representations on some level.

At the core is the argument that the alignment of situation models forms the basis of successful dialogue. These situation models are aligned via primitive and resource-free priming mechanisms. This alignment happens on various levels, such as lexical and syntactic. Next to these primitive processes, there is a repair mechanism to align misaligned representations. Finally, when the primitive processes fail, there are more sophisticated resource-demanding strategies requiring modelling of the interlocutors' mental states.

There are primarily two shortcomings I see in their account. (1) The theory employs a very narrow definition of dialogue, where a face-to-face conversation is the primary form of dialogue that is meant to be studied, but the other forms of dialogue are less interesting derivatives. This also includes automated conversations. Pickering and Garrod (2004) do not discuss the implication of their theory to dialogue automation and the potential of the theory to guide its development. (2) The notion of alignment relies on a notion of successful dialogue which involves the development of aligned representations by the interlocutors. However, such a notion is likely too narrow, as there are dialogue examples in which interlocutors strive for diverging representations, such as in the case of various confrontational dialogues. While the theory lends itself to an account of silence as an indicator of alignment, the authors do not mention it in their account of their theory.

3.6 The interactive stance

The interactive stance is advocated for by Ginzburg (2012). His goal is to develop a theory of conversation rooted in grammar and supported by empirical data from actual conversations. He aims to also cover all different levels of conversation, from something he calls the micro-conversational elements, all the way to macro-level forms. These macro-level forms might include conversations of multiple agents or various dialogue genres. Ginzburg also emphasizes covering various not only different scales but also different stages of dialogue. He divides conversations into the opening , middle and closing stage.

The interactive stance is unique because of its grounding in empirical data. The data to support his theory are primarily focused on language acquisition and computer simulation of language evolution. As a result, he claims that his theory is not leaving out frequently occurring words and construction which have been traditionally left out in approaches based on grammar.

As the name suggests, the theory puts the notion of interaction in the centre. This is further accentuated by the use of evidence from behaviour science about the regular use of language. The strength of this approach is that Ginzberg attempts to connect the formal treatment of dialogue rules with the naturally occurring domain of interactions and the use of grammar.

I find this theory particulary interesting because I have not come across Ginzburg's account of silence and I am not fully sure how it would be accounted for in the theory. It will be likely one of the micro-features which can have significant impacts on the macro level. The specific treatment of silence within this theory, however, remains ambiguous.

3.7 Conclusion

Despite the narrow selection of theories of dialogue, there are some ideas which are relevant to understanding silence and dialogue more broadly. While these ideas might be implicitly used in the development of conversational systems, the explicit awareness of the theoretical frameworks available could further fuel the conceptual and implementational development of these systems. My subsequent exploration of silence in dialogue will be primarily drawing on the concepts from the conversational analysis. Nevertheless, since conversational analysis is not primarily concerned with computational implementation, but rather pragmatic use, there is a potential to widen the scope of the analysis and draw upon other theories presented in this chapter.

Chapter 4

Learning silence

Silence constitutes an important communicative aspect of dialogue. While silence in sociolinguistics and pragmatics has been receiving an increased amount of interest (Jaworski, 1997; Ibrahim and Ambu Muhammad, 2021), the importance of silence has been often neglected in modern dialogue systems. It is common that instead of leveraging the semantically varied use of silence in natural dialogue, dialogue systems implement silence only as a gap constant between dialogue turns. The lack of focus on the diverse use of silence for communication in computer-generated dialogue is thus potentially one of the reasons why modern dialogue systems continue to fall short of the promise of full-fledged human-to-computer dialogue. Implementing a more nuanced use of silence in dialogue systems has the potential, among other things, to make the interaction more human-like and improve the expressive power of the dialogues involved (López Gambino et al., 2019; Adler, 2011; López Gambino et al., 2017).

Recent advances in transformer-based language models, such as GPT-3, trained on large web corpora, offer a new promise for the advancement of human-to-computer dialogue (Vaswani et al., 2017; Brown et al., 2020). GPT-3 has demonstrated substantial gains on many NLP tasks and benchmarks, especially in the context of few-shot performance. GPT-3 has been advertised as a tool that "can

be used to solve virtually any task that involves processing language" (OpenAI, 2022). The following experiment aims to assess the performance of a fine-tuned GPT-3 to model silence in dialogue.

Fine-tuning has been one of the most prevalent approaches to improving the performance of language models. It consists of updating the weights of a pre-trained model by training it on a supervised dataset specific to the desired task (Brown et al., 2020, p.6). The advantage of fine-tuning is the capacity to adapt the model to new benchmarks. However, it is bottlenecked by the need for custom datasets that limit generalization to new tasks. Brown et al. (2020, p.6) identify fine-tuning of GPT-3 as a promising future direction of research.

4.1 Problem formulation

The problem addressed in this experiment can be summarized as *Can silence in dialogue, if encoded as a sequence of characters, be inferred from text alone?* I propose to test the GPT-3 by encoding the silence included in movie subtitle files from the Open Subtitles database as a string of characters. I then fine-tune the GPT-3 model via the OpenAI API with various datasets of different sizes. These datasets consist of dialogue turn pairs filtered from movie subtitles. The goal is to determine the improvement in the performance of a fine-tuned GPT-3 in inferring the length of silence between two subtitle turns when presented with text alone. I begin by testing the model just with an individual movie, then increase to director-level datasets, and then combine two of the director-level datasets to see the overall increase in performance.

To illustrate the problem with an example, consider this conversation turn from the movie Inglourious Basterds by Quentin Tarantino, where a young officer Willi is being convinced to trust him to put down his gun by Lt. Aldo Raine:

Willi: `But But how can I?`

Lt. Aldo Raine: `What choice you got, son?`

What is the length of silence in this turn? After being fine-tuned with a training set from the same movie, the GPT-3 is prompted to infer the silence in this turn. It is fine-tuned to represent the length of the silence as a string of '*'s, so may output the turn in the following form: `But But how can I? ********** What choice you got, son?` In this case, each star represents `10ms` of silence, so the length of the inferred silence is `220ms` which corresponds to `0.22s`.

4.2 Model

The model under investigation is a pre-trained transformer-based language model called GPT-3 (Vaswani et al., 2017; Brown et al., 2020). GPT stands for *Generative Pre-trained Transformer*. Transformers are a class of neural network model using a mechanism of self-attention which allows them to weigh parts of the input to different degrees. The model is trained on a large corpus of web text, resulting in a set of calibrated parameters. In short, the GPT-3 is a 175 billion parameter autoregressive language model known for its in-context learning abilities. Furthermore, by autoregressive, we mean that this model produces an output step-by-step, in such a way that the next input is the output of the previous steps. Hence, in its essence the model is producing outputs from past inputs Vaswani et al. (2017).

Why study the GPT-3? There are other language models, as well as other transformer-based language models, for example, Google's BERT (Devlin et al., 2018). However, GPT-3 is, at the time of writing this thesis, one of the largest transformer-based language models available. Given the track record of GPT-3's success in many NLP tasks, GPT-3 could appear as a promising tool for the task

of modelling silence. If successful, it could be implemented in future dialogue systems to make dialogue pacing more akin to natural dialogue. Although, GPT-3 was not explicitly trained on dialogue data, its few-shot learning ability and its capacity to infer previously unseen aspects of dialogue from its large corpora make it a good candidate to model silence.

The fine-tuning of GPT-3 is accessible via an API by OpenAI. There are different versions of the model, but the following experiment will be using exclusively the latest version, called *text-davinci-002*. The Davinci model is claimed to be "the most capable in the model family and can perform any task the other models can perform and often with less instruction" OpenAI (2022). The fine-tuning and querying of the Davinci model is priced at $0.06 per 1k tokens at the time of writing this thesis.

While the model can be prompted to complete a task just by being given a couple of examples, due to a prompt length limit, fine-tuning is preferable for increasing its performance on a specific task in comparison to the vanilla model (Solaiman and Dennison, 2021). For the best possible result, the estimated number of utterances needed is approximately at least 500 examples (OpenAI, 2022). However, the dataset will enable me to provide more examples as well.

4.3 Experimental set-up

4.3.1 Data

The pre-trained transformer-based language GPT-3 model is fine-tuned on a subset of dialogues from the Open Subtitles dataset (Lison and Tiedemann, 2016). The Open Subtitles database contains over 47100 .txt and .srt subtitle files in English which are more than enough the amount recommended for fine-tuning (OpenAI, 2022).

Fine-tuning is done with pre-selected subsets of the dataset. The initial

selection is done on the level of a movie and a director, where there is a potential for a distinct use of silence in dialogue. Examples of directors who are known for using particular pacing of dialogue in their movies include Quentin Tarantino and Wes Anderson (Braga, 2015). Table 4.1 shows the movies and their respective directors. In total, I have used data from 8 movies, which is a total of 4984 subtitle pairs.

Director	**Movie name**	**code name**	**Movie length**	**# of subtitle pairs**
Quentin Tarantino	Inglourious Basterds	`ing_bast`	02:33:00	1 218
	Pulp Fiction	`pul_fic`	02:34:00	1048
	Django Unchained	`dja_unch`	02:45:00	1506
	Kill Bill 1	`kill_bill`	01:51:00	328
Wes Anderson	Fantastic Mr. Fox	`fan_fox`	01:27:00	-
	Isle of Dogs	`isl_dogs`	01:30:00	-
	The Grand Budapest Hotel	`bud_hot`	01:40:00	-
	Moonrise Kingdom	`moo_kin`	01:35:00	-

Table 4.1: Movie data in the experiment's dataset.

Figure 4.1 shows the silence distribution in the movies that are contained in the dataset. The first four are by Quentin Tarantino, and the last four are by Wes Anderson. Blue areas constitute dialogue turns while their corresponding gaps are the silences in the dialogues. The subtitles are plotted in relation to time which is noted in minutes on the x-axis. One can see that the movies are of different lengths and different silence distributions. For example, Pulp Fiction (`ex1_pulp_fiction`) has subtitles throughout with short or almost no gaps. Contrasting to that is Kill Bill 1 (`ex1_kill_bill_1`) has a sparse dialogue throughout the movie.

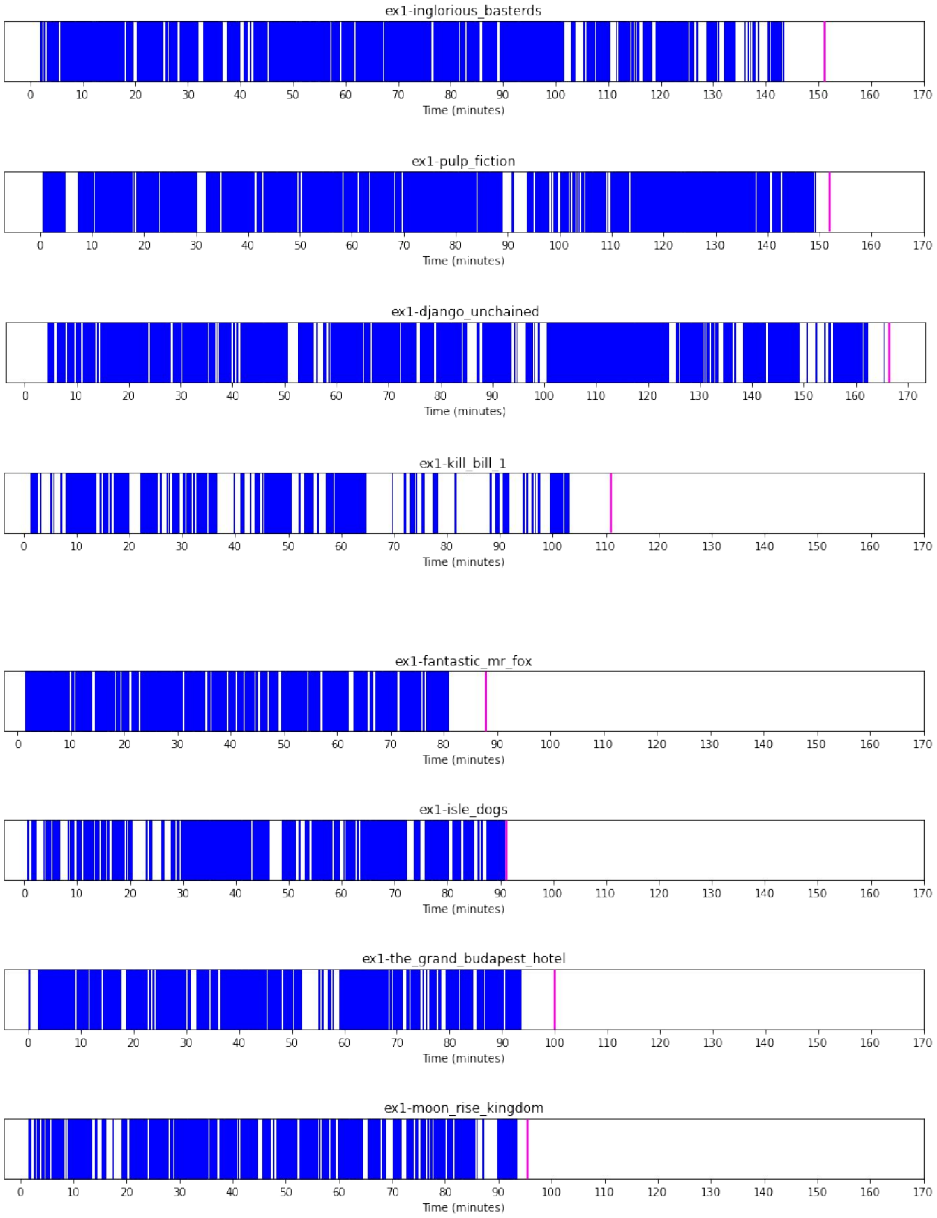

Figure 4.1: Silence length distribution for movies in the dataset; blue areas denote dialogue turns and white areas denote silences; magenta line denotes the end of a movie.

Data pre-processing

The data is pre-processed for fine-tuning to encode the information about the use of silence in dialogue. The subtitle data standardly consists of: (1) a numeric counter indicating the number or position of the subtitle, (2) the start and end time of the subtitle separated by '→', and (3) subtitle text in one or more lines, and (4) a blank line indicating the end of the subtitle. The structure can be generally relied upon to parse a large number of subtitle files.

The data is further filtered for instances when a dialogue turn continues through more than one subtitle. Such subtitles are generally identifiable by beginning and ending with a comma, an ellipsis or no closing character. There are also instances where either a subtitle contains two utterances, each indicated by a dash, or a text which is not a dialogue utterance, often identifiable by being written in all caps. These instances are omitted from the data.

After parsing the data, the challenge is to define basic heuristics for probable dialogue turns. The subtitles that are likely to constitute dialogue turns are then paired. I prioritize pairing of subtitles with their subsequent subtitle if the first subtitle ends with '.', '?', '...', or '!' and the second subtitle begins with a capital letter.

Encoding silence

The corresponding gap for each subtitle pair is calculated in milliseconds. Every 100 milliseconds is represented as a '*'. So, for example, for a two-second gap, the final combination of the subtitle pair is divided with a string of twenty '*'s, such as `Please tell me what you've heard. ********** I've heard nothing.` It is reasonable to assume that the GPT-3 should be good in learning the patterns of a silence encoded as a sequence of one character, as it is generally good in predicting the most likely next token.

The data is then duplicated to form pairs where one string contains the '*'s

and the other string does not, replacing the '*' string with an underscore '_'. This is to form the prompt and completion pair that is used to fine-tune the model.

For example, a training pair might then look like this:

```
prompt:  Are you sure?  _ No.,
completion:  Are you sure?********** No.
```

Training and testing sets

Finally, the data is divided into training and testing sets. The division is done at random, resulting in equally sized datasets, corresponding to half of the original dataset. The training set is used to fine-tune the model and the testing set is used to evaluate the resulting fine-tuned model. The fine-tuning is then further prepared by OpenAI's own data preparation tool which adds necessary dividers and markings for the model and exports the data in the `.jsonl` format.

4.3.2 Fine-tuning

The fine-tuning is done by presenting a prompt and a completion pair, which in this case will be the dialogue pair with (`completion`) and without (`prompt`) the encoded time. When the model is then presented with a different instance of a similar pair, it learns to insert a sequence of '*'s into the completion string. The fine-tuning via the OpenAI API then produces a new model with updated weights based on the training data set.

To evaluate the performance of the model, the model will be then queried with a set of test examples. These examples are subsequently returned with the '*' string which is parsed to a corresponding millisecond value.

4.3.3 Evaluation

I evaluate the performance of the fine-tuned GPT-3 model based on two metrics: (1) the absolute mean error (MAE), and (2) Kullback–Leibler (KL) divergence.

The results of these metrics are then compared to two baselines: (1) the mean baseline, (2) the uniform distribution baseline, and (3) the absolute error baseline.

Mean Absolute Error (MAE)

The MAE is the average absolute difference between the predicted and actual values. The absolute difference is used to account for the fact that the predicted value can be negative. It is calculated as the sum of absolute errors divided by the sample size:

$$MAE = \frac{\sum_{i=1}^{n} |prediction_i - actual_i|}{n} \tag{4.1}$$

The lower the MAE, the better performing the model is in predicting the actual silence length distribution.

Kullback–Leibler (KL) Divergence

The Kullback–Leibler (KL) divergence is a statistical measure of the distance of two probability distributions. It is calculated as the sum of the KL divergences for each dimension:

$$KL(actual, inferred) = \sum_{i=1}^{n} \frac{actual_i}{inferred_i} \log \frac{actual_i}{inferred_i} \tag{4.2}$$

The lower the KL divergence, the smaller difference there is between the actual silence length distribution and the silence length distribution inferred by the model. Hence, the results with lower KL divergence are judged as better performing. The KL divergence of the actual and inferred silence length distribution is then compared to the KL divergence of the two available baselines.

Baselines

I implement two baselines against which I evaluate the model:

1. **Mean baseline:** This baseline consists of the mean silence length of the training dataset. The motivation is to compare the inferred silence length distribution to a scenario where the GPT-3 would have simply learned the mean silence in the training set.

2. **Uniform distribution baseline:** The uniform distribution baseline allows us to compare the result to a scenario where every length of the silence is equally possible. Since the length of the interval $\langle 0, 50\rangle$ corresponds to possible values of the silence length, each corresponding to 10 milliseconds, the probability of any of these values being predicted by the model is $P(1/50) = 0.02$. This probability, where x is possible to silence length can be formally expressed as:

$$P(x) = \begin{cases} \frac{1}{50-0}, x \in [0, 50] \\ 0, x \notin [0, 50] \end{cases} \tag{4.3}$$

While I evaluate the results against both baselines, the mean baseline is more relevant to the absolute mean error (MAE) performance result and the uniform distribution baseline is more relevant to the Kullback–Leibler divergence (KL) performance result.

4.4 Experiments

4.4.1 Movie-level experiment

I begin by evaluating the performance of the GTP-3 model to infer silence when fine-tuned with just one movie. This means that I have only used dialogue data from one movie subtitle file for fine-tuning. To start, I have used the data from the Inglourious Basterds by Quentin Tarantino. I have chosen this film because it is known for its particular use of silence. For example, it features one of the most interesting uses of silence in dialogue in its opening scene where a German

colonel interrogates a French farmer who is hiding innocent Jews below his floor during WWII. Tarantino uses the silence in this instance to build up the tension but also paints the subtleties of the characters (Korenovska, 2017). The following dialogue excerpt shows the deliberate work with silence to build tension, where the silence between the first question and answer goes over to 5 seconds:

Subtitle utterance	**Silence length**
00:17:29,880 → 00:17:33,080 You're sheltering enemies of the state, are you not?	
	00:00:05,120
00:17:38,200 → 00:17:39,240 Yes.	
	00:00:02,800
00:17:42,040 → 00:17:45,800 You're sheltering them underneath your floorboards, aren't you?	
	00:00:03,560
00:17:48,240 → 00:17:49,280 Yes.	
	00:00:02,720
00:17:52,000 → 00:17:54,880 Point out to me the areas where they're hiding.	

Table 4.2: Dialogue turns in the Inglourious Basterds by Quentin Tarantino

After pre-processing based on the aforementioned heuristics, the subtitle file of the movie provides 1218 subtitle pairs. This is then further split into the fine-

tuning and testing sets at random. The fine-tuning set contains 609 pairs and the testing set contains 609 pairs. The fine-tuning for this set cost $3.73.

	GPT-3	**Mean**	**Uniform**
MAE	1.08 s	0.89 s	1.19 s
KL	0.561	2.949	0.654

Table 4.3: Results of the fine-tuning experiment with the Inglourious Basterds by Quentin Tarantino

Figure 4.2 shows the silence length probability distribution of the actual and the inferred silence, as well as the corresponding baselines.

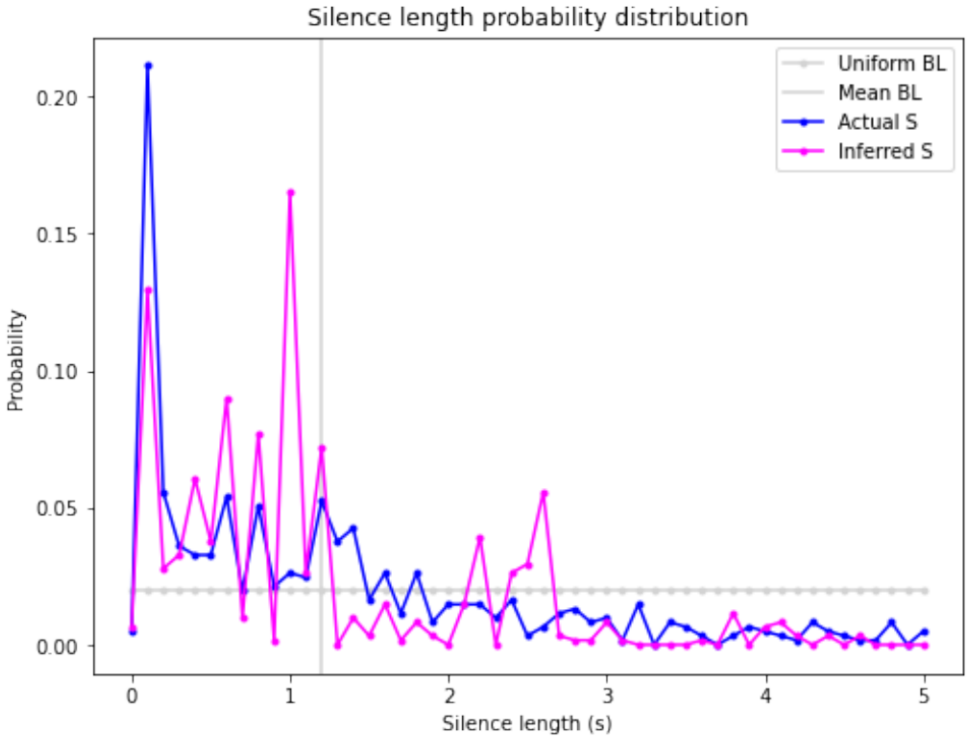

Figure 4.2: Actual and inferred silence distribution in Inglourious Basterds

The mean silence length of the training set was 1.19 seconds. The KL divergence of the inferred silence from the actual silence is 0.561, while the mean baseline

is 2.949 and the uniform distribution baseline is 0.654. If the value of KL was 0, the probability distributions would be identical. The absolute mean error of the GPT-3 is 1.08 seconds while the mean absolute error of the mean baseline is 0.89 seconds and 1.19 seconds for the uniform distribution.

Figure 4.3 shows data for the absolute error for various lengths of silence. The blue dots correspond to the mean baseline MAE values for various silence lengths. It creates a 'V' shape as the lowest error is for the mean and then the predictive power of the baseline decreases by approximately $10ms$ for every $10ms$ change in the predicted silence length, hence the increasing error. It is important to note that in total there are as many data points for the mean baseline MAE as there are data points in the training set. The values averaged to result in the MAE value plotted as a horizontal blue line.

The magenta points in the figure 4.3 correspond to the mean absolute error for the particular silence length. The total MAE is plotted as the horizontal magenta line. The values constituting the mean absolute error per silence length are plotted as grey bars of varying contrast. The absolute error of each subtitle pair was plotted as a line connecting the absolute error value with the mean baseline MAE of the given silence length. The grey bars show the range of error for the particular silence length, as well as an approximate distribution of subtitle pairs per given silence length. The darker the grey, the more subtitle pairs at the particular absolute error range. For example, at the mean (bottom of the 'V' shape) there are relatively many subtitle pairs with an error rate lower than the MAE (darker grey), but still, some have a high absolute error (light grey).

4.4.2 Director-level experiment

To test whether an increased amount of data would improve the GPT-3's capacity to outperform the baselines, I conduct a director-level experiment. In this experiment, I use a dataset consisting of a set of movies from a particular director.

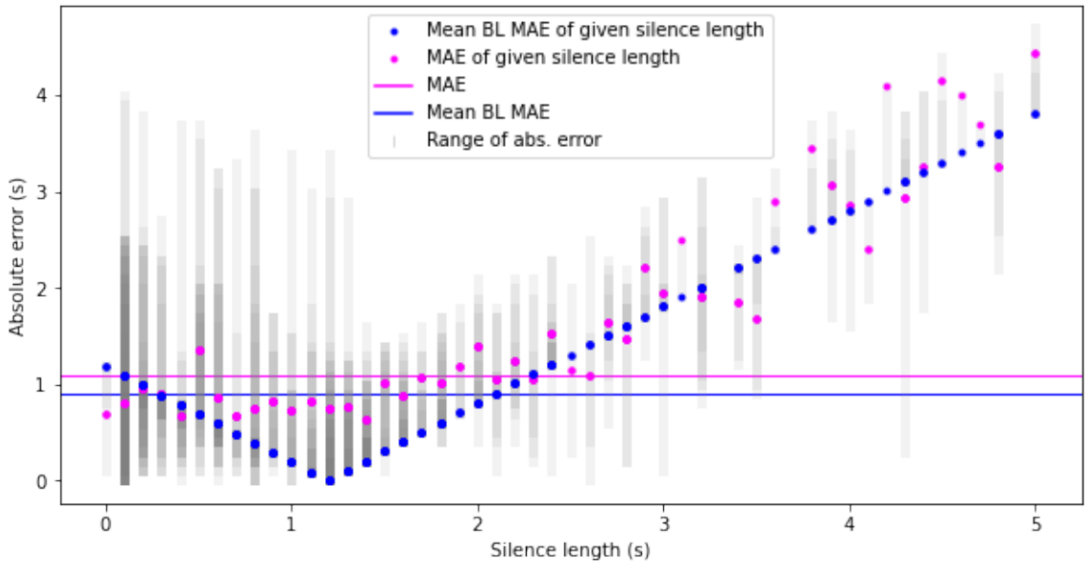

Figure 4.3: Absolute error performance of the Inglourious Basterds by Quentin Tarantino.

Namely, I fine-tune the model with four movies by Quentin Tarantino. The table 4.4 shows the results for the 4 movies:

	GPT-3		Mean		Uniform		
Movie name	**MAE**	**KL**	**MAE**	**KL**	**MAE**	**KL**	**Cost**
Inglourious Basterds	1.08 s	0.561	0.89 s	2.949	1.19 s	0.654	$3.73
Pulp Fiction	1.18 s	0.369	1.26 s	3.924	1.02 s	1.971	$3.79
Django Unchained	0.95 s	0.504	0.81 s	3.453	0.99 s	0.646	$4.17
Kill Bill 1	0.98 s	0.566	0.96 s	3.020	1.22 s	0.899	$0.80

Table 4.4: Movie-level data of 4 movies by Quentin Tarantino

Pulp Fiction constitutes an interesting outlier, as the dialogue in the movie is not featuring a significant amount of silence between turns. This results in the GPT-3 model being able to learn the resulting distribution quite well and significantly outperform the uniform distribution baseline. The probability

distribution of the movie looks as the following:

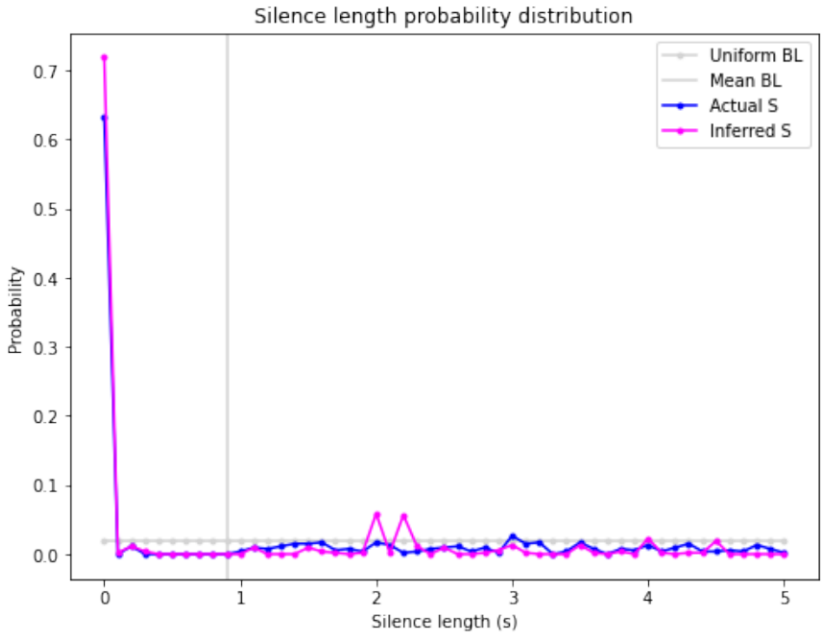

Figure 4.4: Actual and inferred silence distribution for Pulp Fiction by Quentin Tarantino.

The whole aggregate dataset of four movies, after pre-processing, consists of 4108 subtitle pairs. The training cost was $12.43. Table 4.5 shows the results for the aggregate dataset of four movies:

	GPT-3	**Mean**	**Uniform**
MAE	1.27 s	0.94 s	1.13 s
KL	0.426	3.589	0.604

Table 4.5: Results of the fine-tuning experiment with 4 Quentin Tarantino movies

The resulting probability distribution is shown in figure 4.5 with the inferred silence in magenta and the actual silence in blue.

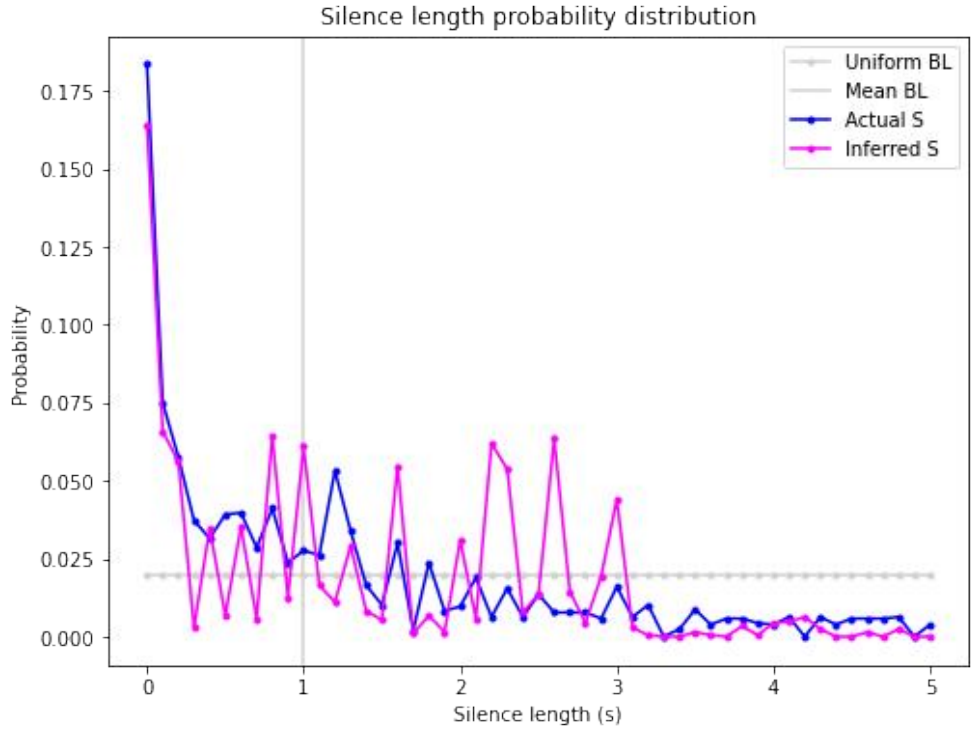

Figure 4.5: Actual and inferred silence distribution for Inglorious Basterds, Django Unchained, Kill Bill 1, and Pulp Fiction by Quentin Tarantino.

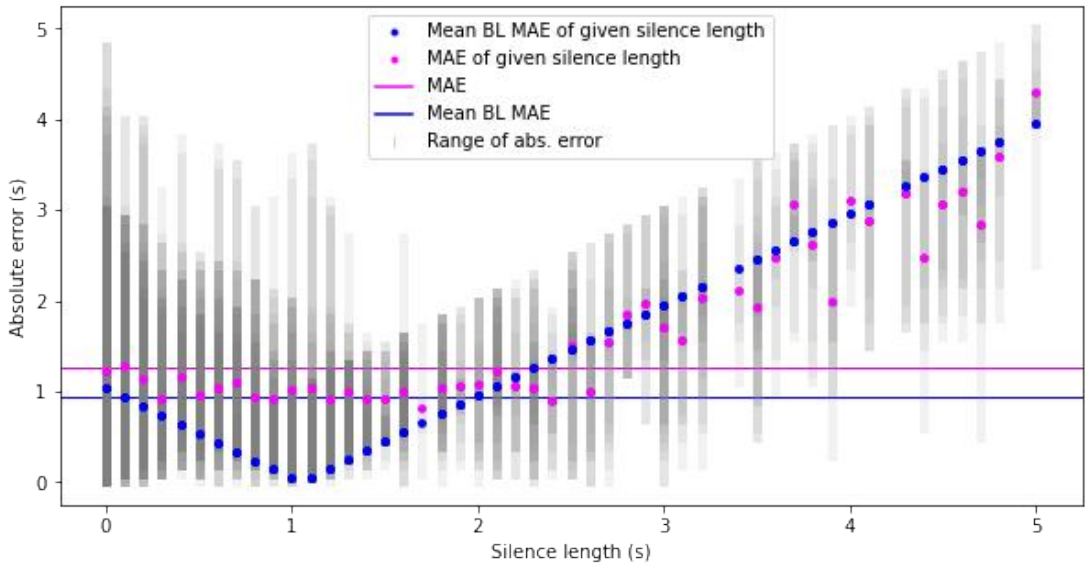

Figure 4.6: Absolute error of director-level experiment with 4 movies by Quentin Tarantino.

The KL divergence of the inferred silence from the actual silence is 0.426 while the KL divergence of the mean baseline is 3.589 and the uniform distribution baseline is 0.604. The absolute mean error of the model is 1.27 seconds while the mean absolute error of the mean baseline is 0.94 seconds and of the uniform distribution is 1.13 seconds. The mean silence length is 1.04 seconds.

Figure 4.6 shows the absolute error per silence length. The interpretation of the plot follows the movie-level result.

4.4.3 Multi-director-level experiment

To see if the trend in improvement of the GPT-3's performance continues with an even larger fine-tuning dataset, I conducted a multi-director-level experiment. Namely, I fine-tune the model with 8 movies from two different directors. I include four movies by Quentin Tarantino and four movies by Wes Anderson. The Quentin Tarantino movies include Inglourious Basterds, Pulp Fiction, Django Unchained, and Kill Bill 1. The Wes Anderson movies include Fantastic Mr. Fox, Isle Dogs, The Grand Budapest Hotel, and Moonrise Kingdom.

In total I have used data from 8 movies, which is a total of 4984 subtitle pairs, amounting to 2492 in the training and testing datasets each. The training cost amounted to $19.72.

	GPT-3	**Mean**	**Uniform**
MAE	1.22 s	0.94 s	0.91 s
KL	0.303	3.912	0.88

Table 4.6: Results of the fine-tuning experiment with 8 movies from 2 directors

Figure 4.7 shows the resulting silence length probability distribution for the actual and inferred silence length for the aggregate dataset of eight movies.

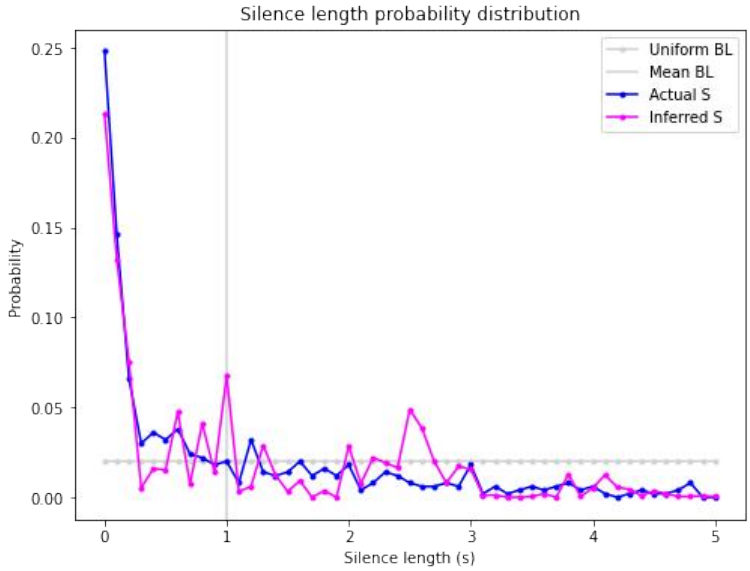

Figure 4.7: Actual and inferred silence for the dataset of movies by Quentin Tarantino and Wes Anderson.

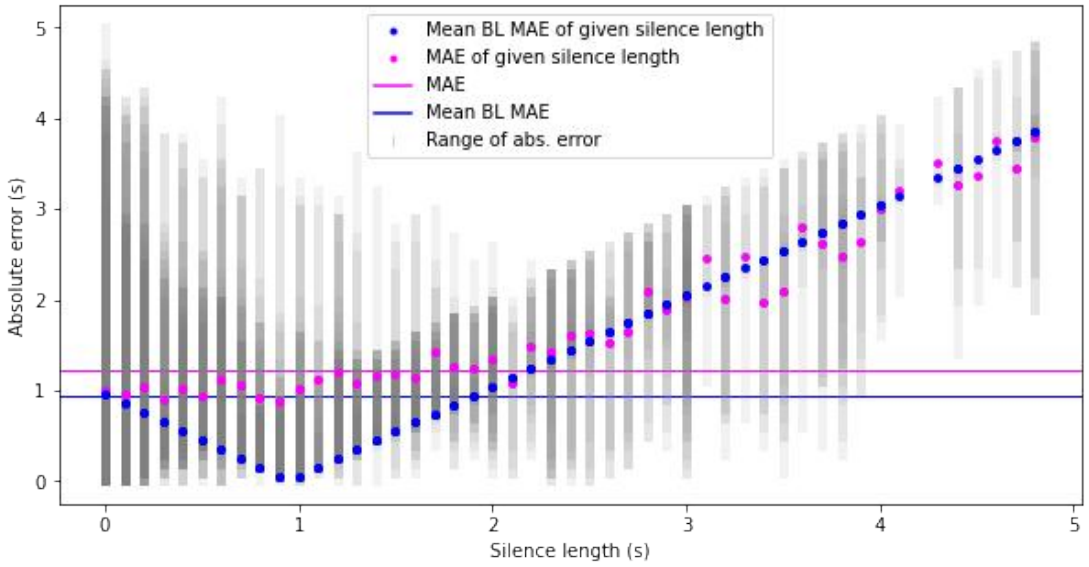

Figure 4.8: Absolute error of multi-director-level experiment with 8 movies.

The KL divergence of the inferred silence from the actual silence is 0.303, while the KL divergence of the mean baseline is 3.912 and the uniform distribution

baseline is 0.88. The absolute mean error of the model is 1.22 seconds while the absolute error of the mean baseline is 0.94 seconds and 0.91 for the uniform distribution. The mean silence length is 1.95 seconds. Figure 4.8 shows the absolute error per silence length. The interpretation of the plot is consistent with the preceding cases.

4.4.4 Impact of fine-tuning dataset size

To evaluate the performance of the GPT-3 model, I conducted a final evaluation. The following figure plots the above-mentioned KL divergence results into a scatter plot. It shows the change in KL divergence with respect to increasing sample size. The trend line indicates that the GPT-3 model gets better in inferring silence as the sample size increases, as the KL divergence decreases with increasing sample size.

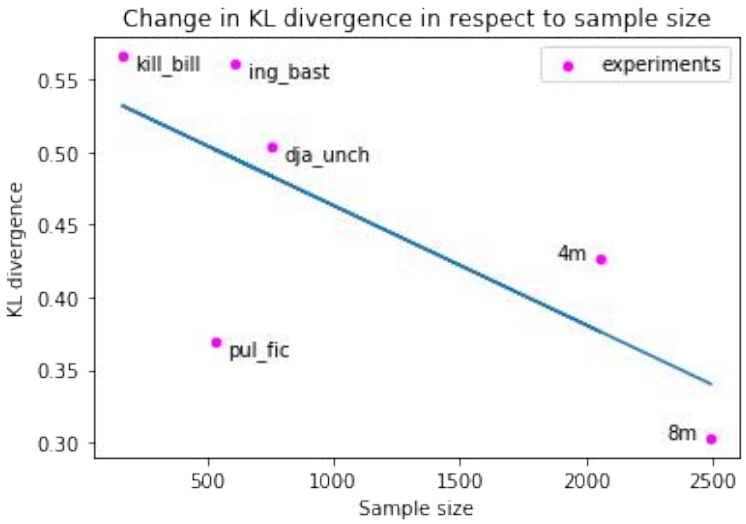

Figure 4.9: Change in KL divergence in respect to sample size.

4.5 Discussion

In the following section, I discuss each experimental set-up individually and then discuss some of the underlying assumptions and remaining questions. Overall, as shown in the progression of the results in the experiments and the figure 4.9, the performance of a fine-tuned GPT-3 in inferring silence length improves with an increasing amount of training data. The increase in performance occurs despite the increased diversity of the movies and directors included. This might mean that the use of silence is not sufficiently diverse across the movies and directors to cause a decrease in predicitive power of the language model due to higher variation. I have also shown that the GPT-3 tends to perform better on KL divergence rather than MAE with smaller datasets. As the dataset size increases, the performance on MAE improves too.

4.5.1 Movie-level result

The movie-level experiment attempted to learn the silence of the Inglorious Basterds by Quentin Tarantino. As seen in the table 4.3, the mean baseline outperforms the fine-tuned GPT-3 on MAE. As shown in figure 4.3, the model struggles to learn the silence distribution in the long silences, where the MAE averages per silence length are often higher than the mean baseline MAE by more than a second. Nevertheless, in some cases, the model also succeeds in outperforming the baseline. Further analysis could therefore focus on evaluating to what extent these performance gains in smaller datasets are a matter of chance or whether there are properties of the data that allow the model to learn the silence from fewer data.

The model performs better than the uniform distribution baseline when it comes to KL. One can notice significant spikes in figure 4.3 around the mean at around one second. It is thus plausible that the model in this case attempts to approximate the silence length means as a possible strategy.

4.5.2 Director-level results

The director-level experiment focuses on the analysis of four movies by Quentin Tarantino. First, it is interesting to pay brief attention to the individual results of the movies that constitute the dataset. Table 4.4 list the results for the four movies. One can immediately notice the differences that arise from the different subtitle datasets, such as wildly different silence length means. While these are movies from the same director, Inglorious Basterds and Kill Bill 1 feature quite different dialogue structures. Kill Bill 1 has a very low silence length mean, indicating that it is not a very dialogue-heavy movie. This highlights some of the limitations that might come into play when using subtitle datasets to model silence.

In table 4.5, the results indicate that the mean baseline performs better than the fine-tuned GPT-3 on MAE. The absolute error of the model is 1.27 seconds while the absolute error of the mean baseline is 0.94 seconds. Nevertheless, the model outperforms the uniform distribution baseline on KL. As it can be seen both in the figure 4.5 and 4.6, subtitle turns with silence length below 1.5 seconds are constituting a majority of the dataset. Exploiting this statistical feature of the dataset is perhaps one of the main reasons why the model is able to score better than the uniform baseline on the KL divergence.

4.5.3 Multi-director-level results

The multi-director-level experiment focused on the analysis of eight movies by Quentin Tarantino and Wes Anderson. As shown in the table 4.6, the GPT-3 outperforms the mean baseline on MAE and KL. The absolute error of the model is 1.22 seconds while the absolute error of the mean baseline is 0.94 seconds and 0.91 seconds for the uniform distribution. The model outperforms the uniform distribution and mean baseline in terms of KL.

While in figure 4.7, there is a visible small spike at the mean at 0.95 seconds

the model seems to be successful in inferring the overall silence length rather than just learning the mean. It likely again exploits the notable disproportionality of short tuns in comparison to long silences.

Figure 4.8 shows that with an increasing number and more consistent distribution of silence lengths throughout the dataset, the model is also becoming better at learning longer silences, as the magenta points follow closer to the mean baseline.

4.5.4 Remaining questions

The question remains whether the improved performance is a product of GPT-3's capacity to be fine-tuned effectively in instances of high-frequency of no silence and short silences, such as in the case of Pulp Fiction. The percentage of these short turns increases, as more movies are included. In particular, some of the movies contain disproportionaly short turns, skewing the the mean silence length even further. This might be further tested by constructing datasets specifically with a high variation in silence length across the available range.

There is a series of assumptions that underpin the experiment. Firstly, one should ask to what extent it is justified to reduce the varied uses of silence in dialogue to a uniform string of one character, such as '*'s. One can argue that such an approach cannot capture the nuances of the use of silence, as there are various uses of silence in natural language. This might require use different type of charcters to represent different types of silence or perhaps resort to a different approach to encoding altogether. (Kurzon, 2007) proposes a typology of silence that could be used for this purpose in the future.

Secondly, some limitations come along with the use of the subtitle dataset. Mainly, subtitles do not necessarily map on natural use of silence in dialogue as such, but rather optimize for convenient legibility (Bannon, 2010). This might mean that their resolution is not as great when it comes to short, the more

condensed dialogue turns, as the subtitle would be spread out to maximise time for reading. Furthermore, subtitles do not allow us to determine whether silence in dialogue is filled with a particular activity or some events in the movie. This might cause a fraction of the dataset used in the experiment might not meet some of the stricter definitions of a dialogue dataset. Boundary detection in subtitles has been, for example, studied by Donabauer et al. (2021) and could be implemented in future iterations of this study to improve the quality of the dataset.

Finally, the assumption that it is possible to predict inter-turn silence only from the text of the previous turn, rather than a longer dialogue context, is a strong simplification. The reason for not having experimented with more context is that it requires the processing of a larger dataset, where the preceding context meets the condition for continuous dialogue, set in this experiment to maximum silence of 5 seconds. Thus, in the datasets that I have worked with, there would be only a few turns which could include a sufficiently long context preceding the turn.

4.6 Future work

As mentioned, a natural progression of the experiment would be an experiment with a larger portion of subtitle data which would allow for the inclusion of more than one turn and silence between several previous turns to make a prediction. This could provide an interesting avenue for the discussion of the role of context in the inference of silence length and its impact on the performance of language models, such as GPT-3.

Another interesting development could be using sentiment analysis to detect different types of silence and experiment with the capacity of language models to infer different types of silence accordingly. This could, for example, follow the approach proposed by (Shi and Yu, 2018) who integrate multimodal information

(acoustic, dialogic and textual) in the model. Similar data could be sourced from the movies themselves in addition to the subtitles. Movie audio tracks could be also used to improve precision of the boundary detection for individual dialogue turns.

Future work might include experimentation with an even larger dataset and comparing the performance of various language models. It would be interesting to explore whether and where the upper bound on the improvement of the GPT-3 when it comes to data it was not explicitly trained on. If one would be truly serious about machine-to-human dialogue, it appears necessary to begin to build models which can account for silence. Such an endeavour lends itself to the question of what datasets and models are most suitable for such a task.

4.6.1 Human evaluation

Given the nuanced nature of the use of silence in dialogue, a human evaluation experiment would be relevant to establish a benchmark to be compared against the performance of computational models. Especially in the instance when computational models aim to approximate a particular human performance, collecting human evaluation can serve as a useful way of establishing a point of reference. Furthermore, it can be used to mitigate the shortcomings of automatic evaluation which does not always correlate with human judgement. In their work, van der Lee et al. (2019) argue that one should conduct a human evaluation of a machine-generated dialogue whenever possible. In the following section, I cover some references that are informative when it comes to a human evaluation study design and implementation. I also outline a possible starting point for a human evaluation study of silence in dialogue.

Challenges of human evaluation

Human evaluation of computer-generated dialogue is most commonly conducted via online surveys. Hämäläinen and Alnajjar (2021) claim that the most frequently utilised strategy is to ask evaluators to rate outputs of a model on a scale from 1 to 5, also known as the Likert scale. Although the Likert scale might be the most common evaluation method, Novikova et al. (2018) argues that continuous scales have been shown to give more nuanced evaluations and were preferred by evaluators. As Santhanam and Shaikh (2019) show, it is therefore important to remain cognisant of the evaluation task design and presentation, as it can easily affect the consistency and quality of human judgements.

While human evaluation can be reliable, there has been a large variation among the criteria employed in different studies. As Belz et al. (2020) find, the inter-evaluator agreement and self-consistency tend to be low. This variation results in inconsistencies, lack of replicability, and generalisability (Howcroft et al., 2020; Hämäläinen and Alnajjar, 2021; Celikyilmaz et al., 2021). To mitigate this issue, Howcroft et al. (2020) argue that it is important to use consistent evaluation criteria across different studies and to this end proposes a classification system with the goal of increased reproducibility. Finally, as Hämäläinen and Alnajjar (2021) stresses, it is important to collect demographic evaluators to avoid bias.

Human evaluation design

A human evaluation in the context of the problem of inferring silence length would serve two primary points: (1) to evaluate how dow humans perform on the task of inferring silence between two dialogue turns, as presented to the GPT-3 language model, and (2) to generate a training dataset of natural use of silence in dialogue that could potentially overcome some of the limitations of the use of subtitles. By the shortcoming of subtitles, I primarily mean that they do not perfectly map on the natural use, as their timestamps do not perfectly match when the movie

characters speak.

There are various ways in which silence could be presented to the evaluators, as it is perhaps not the most natural to think of silence in terms of a sequence of '*'s. One way could be to visualise silence as a distance between two consecutive dialogue turns. The survey then should allow the user to drag and drop the utterance further or closer to represent the length of the corresponding silence. Another way could be to animate the survey interface, so the time of the utterances appearing would map onto the silence in the dialogue. Finally, one could use text-to-speech functionality, so users can hear the length of silence used in the turn to evaluate its appropriateness. The results would be then evaluated by calculating the difference in the length of silence inferred by the language model.

To enable reproducibility of such a study, it could be based on a criteria classification system, such as the one by Howcroft et al. (2020). In the case of Howcroft et al. (2020), the evaluation would fall under *the goodness of output relative to external frame of reference*. Namely, it could probably best utilise the criteria of naturalness, mainly in terms of the form silence is used in the output. Howcroft et al. (2020) see naturalness as synonymous with clarity and human-likeness. Another criterion for consideration would be appropriateness. However, this criterion is defined as more context-dependent.

4.6.2 Incremental speech generation

Future work utilising the results of this work could extend the application to the problem of incremental speech generation. The problem of inferring time gaps between dialogue turns has potential application to incremental speech generation. Incremental speech generation aims to overcome the limitation of conversational systems where the output needs to be processed before a response can be generated. This produces unnatural gaps of silence which are often perceived negatively by its users. For example, Skantze and Hjalmarsson (2013)

propose a conversational system that incrementally interprets spoken input, while simultaneously plans, realises and self-monitors the system response. The incremental version has a shorter response time and is perceived as more efficient by the users. Other benefits of well-timed responses include increasing users' perception of humanness and social presence, but also leading to greater satisfaction with the overall interaction (Gnewuch et al., 2018).

While one strategy is to simply decrease the response delay, such as presented by Tsai et al. (2019), the other option is to leverage the naturally occurring use of silence in dialogue to enable sufficient processing time. The use of silence, however, has to be implemented in a way so the use of silence resembles the use in natural dialogue or it has to be complemented with filler words, such as 'umm' or 'uhh' (López Gambino et al., 2019, 2017; Betz et al., 2018). Another option has been proposed to adapt behaviour to the perceived cognitive load created by the conversation (Lopes et al., 2018). Thus, understanding when to use silence and to use it only in the right amount is paramount to achieving a successful incremental speech generation.

4.7 Overview

I have presented an experiment to test the capacity of GPT-3 to infer the length of silence in dialogue. The experiment was conducted on a dataset consisting of a set of movies from two directors, namely Quentin Tarantino and Wes Anderson. I have shown that GPT-3 can infer silence length from a pure text when fine-tuned on a dataset with silence encoded as a sequence of characters and that the efficiency of the fine-tuning process is improved with the increased size of the dataset despite increased diversity of the dataset.

Chapter 5

Conclusion

I began by mapping out the historical development of the study of dialogue, followed by an overview of the most prominent theories of dialogue. The overview shows that the research on general theories of dialogue remains sparse and scattered across different disciplines. I have discussed some of the treatments of silence in dialogue. Overall, silence is understudied and not significantly covered as a component of dialogue. The theory overview also highlights a notable divide between the study of dialogue as human behaviour and the theories used to study computer-generated dialogue. Thus, this work shows that there is a room for dialogue theorists to engage with computational methods and for researchers working on computational dialogue systems to engage with dialogue theories in order to develop a unified understanding of dialogue.

In the second part of the thesis, I tested the performance of the fine-tuned GPT-3 language model on inferring silence in dialogue turns. The results show that the performance of the model improves with the increasing size of the dataset despite its increased variation in terms of different movies and directors. This means that there were no major differences in the use of silence by the utilised movies and within the styles of the two different directors. The result also showcases different strategies taken by the GPT-3 respective to the particular

dataset. I have proposed an approach to subtitle pre-processing, as well as encoding of silence for fine-tuning that leverages GPT-3 capacity for the next character prediction.

Overall, the most notable contribution of this work includes the overview of available theories of dialogue and a framework for the utilisation of timestamped datasets, such as subtitles, for the fine-tuning of transformer-based language models, such as the GPT-3, to improve the pacing of computer-generated dialogue. The work shows that while GPT-3 might appear as a powerful language model, a successful emulation of an aspect of dialogue, such as silence, requires also reliable fine-tuning datasets and evaluation strategies.

5.1 Limitations

Theories of dialogue

The theories of dialogue explored in this work are limited to mostly western-centric analytical tradition. This should not be by no means considered exhaustive, as it leaves out many other socio-cultural approaches to understanding dialogue (Wierzbicka, 2006). While not included in this study, other theories and cultural approaches to dialogue should be also considered.

Learning silence

The experimental set-up relies on a series of assumptions that do not reflect how silence is processed in natural dialogue. Most notably, the experiment assumes that it is possible to infer the length of silence only from the text of the previous utterance, while in reality, humans need an extensive understanding of context. Moreover, it assumes that silence is reduced to one type of a string, while there are various types and uses of silence in natural dialogue.

5.2 Ethics statement

Silence as silencing

The study of silence should not and cannot be separated from the use of silence in dialogue, where particular uses of silence in dialogue have been forcefully used to silence specific (minority) groups in society based on their gender, race, or other characteristics. Some of the strategies of silencing in society include ridicule, enforcement of family hierarchies, male-controlled media, anti-woman educational policies, making women's bodies political battlegrounds, censorship, racism, homophobia, and terrorism (Houston and Kramarae, 1991). There is a risk of these patterns becoming reflected in dialogue systems through the training data and/or perpetuated by dominant means of studying silence.

Bias in subtitles

Movie subtitles are subject to existing societal biases, such as male characters being often more frequently represented as protagonists while leading characters represented by women and/or racial/ethnic minorities are less prominent (Erigha, 2015). This might be skewing the training dataset to represent the use of silence of dialogue by a particular group of people. Such a distortion might contribute to the perpetuation of harmful social norms and prejudices and should be actively addressed in future research.

Social impacts of large language models

There are inherent risks to the wide avaliability of large language models, such as GPT-3. With the increasing availability of large language models, such as GPT-3, there is a risk of the models being used for harmful purposes, such as the generation of fake news, hate speech, and other forms of misinformation. The use of large language models for such purposes should be actively discouraged and

prevented (McGuffie and Newhouse, 2020).

Moreover, training large models, such as GPT-3 requires vast amounts of computing power that results in a significant carbon footprint. This impact does not only occur at the training, but at the fine-tuning and querrying stages as well. Steps can be taken to minimise the carbon footpring, which also incudes an effective consumption tracking and incentivization of responsible research (Henderson et al., 2020).

Bibliography

Adler, S. (2011). Silence in the graphic novel. *Journal of Pragmatics 43*(9), 2278–2285.

Anderson, R., L. A. Baxter, and K. N. Cissna (2003). *Dialogue: Theorizing difference in communication studies*. SAGE publications.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Austin, J. L. and G. J. Warnock (1962). *Sense and sensibilia*, Volume 83. Clarendon Press Oxford.

Bakhtin, M. (2010). *The dialogic imagination: Four essays*. University of texas Press.

Bannon, D. (2010). *The Elements of Subtitles, Revised and Expanded Edition: A Practical Guide to the Art of Dialogue, Character, Context, Tone and Style in Subtitling*. Lulu. com.

Belz, A., S. Mille, and D. M. Howcroft (2020, December). Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, Dublin, Ireland, pp. 183–194. Association for Computational Linguistics.

Betz, S., B. Carlmeyer, P. Wagner, and B. Wrede (2018). Interactive

hesitation synthesis: modelling and evaluation. *Multimodal Technologies and Interaction 2*(1), 9.

Bohm, D., P. M. Senge, and L. Nichol (2004). *On dialogue*. Routledge.

Braga, P. (2015). *Words in action. Forms and techniques of film dialogue*. Peter Lang.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Bruneau, T. J. (1973). Communicative silences: Forms and functions. *Journal of communication 23*(1), 17–46.

Celikyilmaz, A., E. Clark, and J. Gao (2021, May). Evaluation of Text Generation: A Survey. *arXiv:2006.14799 [cs]* (0), 0. arXiv: 2006.14799.

Clark, H. H. (1996, May). *Using Language* (1st edition ed.). Cambridge England ; New York: Cambridge University Press.

Clerbout, N. and Z. McConaughey (2022). Dialogical Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.)., pp. 0. Metaphysics Research Lab, Stanford University.

Cuayáhuitl, H., S. Renals, O. Lemon, and H. Shimodaira (2005). Human-computer dialogue simulation using hidden markov models. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pp. 290–295. IEEE.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Donabauer, G., U. Kruschwitz, and D. Corney (2021, April). Making Sense of Subtitles: Sentence Boundary Detection and Speaker Change Detection in Unpunctuated Texts. In *Companion Proceedings of the Web Conference 2021*, WWW '21, New York, NY, USA, pp. 357–362. Association for Computing Machinery.

Duhoe, A. A. A. and E. Giddi (2020, May). Semantic Analysis of Silence in Conversational Discourse. *Journal of Linguistics and Foreign Languages 1*(1), 18–32. Number: 1.

Ephratt, M. (2011). Linguistic, paralinguistic and extralinguistic speech and silence. *Journal of pragmatics 43*(9), 2286–2307.

Erigha, M. (2015). Race, gender, hollywood: Representation in cultural production and digital media's potential for change. *Sociology compass 9*(1), 78–89.

Floridi, L. and M. Chiriatti (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines 30*(4), 681–694.

Frampton, M. and O. Lemon (2005). Reinforcement learning of dialogue strategies using the user's last dialogue act. In *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp. 0.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language* (0), 27–52.

Ginzburg, J. (2012). *The interactive stance: Meaning for conversation*. Oxford University Press.

Gnewuch, U., S. Morana, M. T. P. Adam, and A. Maedche (2018). Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *26th European Conference on Information*

Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018. Ed.: U. Frank, pp. 143975.

Goldhill, S. et al. (2008). *The end of dialogue in antiquity*. Cambridge University Press.

Goodson, I. and S. Gill (2014). *Critical narrative as pedagogy*. Bloomsbury Publishing USA.

Gregory, V. (1983). The socratic elenchus. *Oxford Studies in Ancient Philosophy 1*.

Grice, P. (1989a). *Studies in the Way of Words*. Harvard University Press.

Grice, P. (1989b). *Studies in the Way of Words*. Harvard University Press.

Henderson, P., J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research 21*(248), 1–43.

Hämäläinen, M. and K. Alnajjar (2021, August). Human Evaluation of Creative NLG Systems: An Interdisciplinary Survey on Recent Papers. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, Online, pp. 84–95. Association for Computational Linguistics.

Houston, M. and C. Kramarae (1991). Speaking from silence: Methods of silencing and of resistance. *Discourse & Society 2*(4), 387–399.

Howcroft, D. M., A. Belz, M.-A. Clinciu, D. Gkatzia, S. A. Hasan, S. Mahamood, S. Mille, E. van Miltenburg, S. Santhanam, and V. Rieser (2020, December). Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, Dublin, Ireland, pp. 169–182. Association for Computational Linguistics.

Ibrahim, B. and U. Ambu Muhammad (2021, July). The most Powerful Thing You'd Say Is Nothing at all: The Power of Silence in Conversation. In *Types of Nonverbal Communication [Working Title]*. IntechOpen.

Ibrahim, B. and U. A. Muhammad (2021). The most powerful thing you'd say is nothing at all: The power of silence in conversation. *Types of Nonverbal Communication*, 125.

Jaworski, A. (1997). *Silence: Interdisciplinary Perspectives*. Walter de Gruyter. Google-Books-ID: OM4ueFfoRfcC.

Junmei, Z. and L. William (2019). Predicting customer call intent by analyzing phone call transcripts based on cnn for multi-class classification. In *8th International Conference on Soft Computing, Artificial Intelligence and Applications*, pp. 9–20.

Komeili, M., K. Shuster, and J. Weston (2022, May). Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 8460–8478. Association for Computational Linguistics.

Korenovska, L. (2017). Language game as means of persuasion (basing on the film inglourious basterds by quentin tarantino). *Problems of modern literary studies* (25), 35–46.

Kurzon, D. (2007, October). Towards a typology of silence. *Journal of Pragmatics 39*(10), 1673–1688.

Lee, C., S. Jung, S. Kim, and G. G. Lee (2009). Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication 51*(5), 466–484.

Lestary, A., N. Krismanti, and Y. Hermaniar (2018a). Interruptions and silences in conversations: a turn-taking analysis. *PAROLE: Journal of Linguistics and Education 7*(2), 53–64.

Lestary, A., N. Krismanti, and Y. Hermaniar (2018b, October). Interruptions and Silences in Conversations: A Turn-Taking Analysis. *PAROLE: Journal of Linguistics and Education 7*, 64.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Linell, P. (1998). *Approaching dialogue: Talk, interaction and contexts in dialogical perspectives*, Volume 3. John Benjamins Publishing.

Lison, P. and J. Tiedemann (2016, May). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 923–929. European Language Resources Association (ELRA).

Lopes, J., K. Lohan, and H. Hastie (2018). Symptoms of cognitive load in interactions with a dialogue system. In *Proceedings of the workshop on modeling cognitive processes from multimodal data*, pp. 1–5.

Lorenc, P. (2021). Joint model for intent and entity recognition. *arXiv preprint arXiv:2109.03221* (0), 0.

López Gambino, S., S. Zarrieß, and D. Schlangen (2017, August). Beyond On-hold Messages: Conversational Time-buying in Task-oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbrücken, Germany, pp. 241–246. Association for Computational Linguistics.

López Gambino, S., S. Zarrieß, and D. Schlangen (2019). Testing Strategies For Bridging Time-To-Content In Spoken Dialogue Systems. In L. F. D'Haro, R. E.

Banchs, and H. Li (Eds.), *9th International Workshop on Spoken Dialogue System Technology*, Lecture Notes in Electrical Engineering, Singapore, pp. 103–109. Springer.

McGuffie, K. and A. Newhouse (2020). The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

McTear, M. (2020, October). Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. *Synthesis Lectures on Human Language Technologies 13*(3), 1–251.

Meng, H. M., C. Wai, and R. Pieraccini (2003). The use of belief networks for mixed-initiative dialog modeling. *IEEE Transactions on Speech and Audio Processing 11*(6), 757–773.

Merriam-Webster (2022). Definition of DIALOGUE.

Moor, J. H. (2001). The status and future of the turing test. *Minds and Machines 11*(1), 77–93.

Murphy, J. G. (1998). Kant on theory and practice. In *Character, Liberty, and Law*, pp. 5–32. Springer.

Nadeau, D. and S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*(1), 3–26.

Novikova, J., O. Dušek, and V. Rieser (2018, June). RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, pp. 72–78. Association for Computational Linguistics.

OpenAI (2022). OpenAI API.

Pangaro, P. (2017). Questions for conversation theory or conversation theory in one hour. *Kybernetes* (0), 0.

Pask, G. (1975). Conversation, Cognition and Learning. In *Amsterdam, Elsevier*. CUMINCAD.

Pask, G. (1976). Conversation theory. *Applications in Education and Epistemology* (0), 0.

Pickering, M. J. and S. Garrod (2004, April). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences 27*(2), 212–225. Publisher: Cambridge University Press.

Potts, C. (2009). Formal pragmatics. *The Routledge Encyclopedia of Pragmatics* (0), 167–170.

Sacks, H., E. A. Schegloff, and G. Jefferson (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pp. 7–55. Elsevier.

Santhanam, S. and S. Shaikh (2019, October). Towards Best Experiment Design for Evaluating Dialogue System Output. In *Proceedings of the 12th International Conference on Natural Language Generation*, Tokyo, Japan, pp. 88–94. Association for Computational Linguistics.

Scheffler, K. and S. Young (2002). Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of HLT*, Volume 2, pp. 0.

Searle, J. R. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), *Language, Mind and Knowledge*, pp. 344–369. University of Minnesota Press.

Shi, W. and Z. Yu (2018, July). Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 1509–1519. Association for Computational Linguistics.

Skantze, G. and A. Hjalmarsson (2013, January). Towards incremental speech generation in conversational systems. *Computer Speech & Language 27*(1), 243–262.

Solaiman, I. and C. Dennison (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems 34*.

Stivers, T., N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon, et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences 106*(26), 10587–10592.

Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems 27*.

Taylor, A., M. Marcus, and B. Santorini (2003). The penn treebank: an overview. *Treebanks* (0), 5–22.

Tsai, V., T. Baumann, F. Pecune, and J. Cassell (2019). Faster Responses Are Better Responses: Introducing Incrementality into Sociable Virtual Personal Assistants. In L. F. D'Haro, R. E. Banchs, and H. Li (Eds.), *9th International Workshop on Spoken Dialogue System Technology*, Lecture Notes in Electrical Engineering, Singapore, pp. 111–118. Springer.

Turing, A. M. (1950, October). Computing Machinery and Intelligence. *Mind LIX*(236), 433–460.

van der Lee, C., A. Gatt, E. van Miltenburg, S. Wubben, and E. Krahmer (2019). Best practices for the human evaluation of automatically generated

text. In *Proceedings of the 12th International Conference on Natural Language Generation*, Tokyo, Japan, pp. 355–368. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*(1), 36–45.

Wierzbicka, A. (2006). The concept of 'dialogue'in cross-linguistic and cross-cultural perspective. *Discourse Studies 8*(5), 675–703.