Where Do We Go Now? An Overview of Conversational A.I.

December 13, 2021

David C. Morley

Abstract

The field of conversational A.I. is explored with particular interest to its historical background, shortcomings, and the future of task-oriented systems. Recent approaches are highlighted and serve as a starting point for future research and development. Lastly, proposals for future studies to build on the current successes of the field are outlined.

1. Introduction

In 1950, Alan Turing, one of the fathers of modern computing, stated his belief that machines would be conversational by the year 2000 (Turing, 1950). However, time has shown that conversational A.I. is much more difficult to create than initially thought. Over a decade Turing's prediction, one of the first machines capable of taking this test, ELIZA (Weizenbaum, 1966), caught the public eye. ELIZA[1] dazzled users by searching for keywords in a given input and applying an associated rule to generate a response. It also could remember previous states that it could draw upon if no keywords were found. Lastly, Weizenbaum created a policy to rank the importance of keywords if multiple appeared in a given input.

Along with being historically interesting, ELIZA also serves as a representative framework for conversational systems built today (Neff and Nagy, 2016). Though the names of the components vary, nearly all conversational systems contain the same components. Concretely, these are Natural Language Understanding, Dialogue Management/Belief Tracking, Policy, and Natural Language Generation (see, for

[1] ELIZA was named after Eliza Doolittle, the protagonist of both *Pygmalion* and *My Fair Lady*, because it appeared civilized despite its simple background

example, Bobrow et al., 1977; Rambow and Korelsky, 1992; Wen et al., 2017; Lubis et al., 2020). As computational models based on statistical methods have increased in ability, reliance on ELIZA's string matching and lookup table methods have decreased. However, the guaranteed reproducibility of the rule-based approach makes it preferred in applications where control over a system's output is essential. At the other end of the spectrum, advancements in neural networks have enabled researchers to eliminate the need for separate model components to favor an end-to-end approach. Unfortunately, this shortcut in engineering comes with other shortcomings in performance. All conversational systems will fall somewhere in the spectrum between the purely rule-based approach of ELIZA and the pure end-to-end approach.

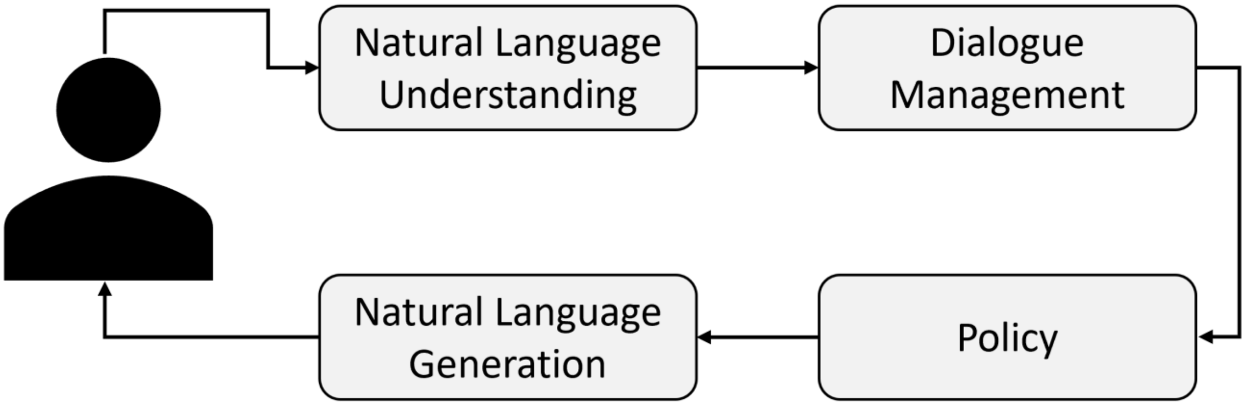

Figure 1: A diagram of the standard modules of a conversational system.

In this review, I give a more thorough overview of conversational A.I. by building upon the discussion of types of systems created. While systems are often referred to in terms of how they fall in the spectrum of rule-based and end-to-end, a more relevant distinction could be the purpose of their creation. Task-oriented systems are designed to help the user toward a specific goal and are very limited in scope, while open-domain systems are intended to discuss any topic. Using the knowledge of the types and purposes of conversational systems, I will discuss the general limitations of present methods and models of both types of systems and in the field as a whole. This will motivate a discussion of recent promising approaches in task-oriented conversational systems that I believe respond to these current needs. Lastly, I will propose areas for future research and model design to build on the strengths of the promising research.

The structure of the paper is as follows. First, Section 2 gives an overview of conversational A.I. with special attention given to current shortcomings. Section 3 builds on the previous section by illustrating selected successful methods. In Section 4, I detail proposals for future research and experimentation. Lastly, Section 5 provides a conclusion to tie a bow around the more salient arguments of this work.

2. Background

As mentioned in the introduction, historically, there has been a significant split in the field of conversational A.I. between rule-based and end-to-end systems (Zhang et al., 2020). However, statistical methods have been used to augment each aspect of the rule-based systems making this bifurcation less distinctive (Dai et al., 2020). Current research tends to fall into two camps: task-oriented vs. non-task-oriented systems. Though both task-oriented and non-task-oriented systems can be either rule-based or end-to-end, task-oriented systems are usually more rule-based, while non-task-oriented are more end-to-end.

2.1 Task-Oriented vs. Non-Task-Oriented

Task-oriented (also called goal-oriented) dialogue systems are designed to assist the user in completing a task on a specific domain (Jurafsky and Martin, 2021). To ensure success in a task, these systems are commonly built around a hand-crafted ontology (Lubis et al., 2020). An ontology is a knowledge base that provides the types of utterances that can be understood and responded to by the system (Zhang et al., 2020). These ontologies can take many forms but commonly contain scripted outputs for a closed set of understandable inputs. For this reason, these systems tend to lean toward the rule-based side of the spectrum but often use statistical methods for some components. These systems are distinguished from non-task-oriented systems, whose primary function is to engage the user in conversation about most any subject.

Non-task-oriented systems, also commonly called open-domain systems or chatbots, can be conversational on many topics but have no guiding purpose. Their main goal is to mimic the unstructured conversations of informal human communication (Jurafsky and Martin, 2021). They tend to jump around different topics and often end in loops or absurd utterances. Additionally, they tend to try so hard to sound human-like while missing the higher-level goals of conversational flow (Lubis et al., 2020). These shortcomings aside, chatbots are far more convenient to create because they skip the ontology for a more data-driven approach (see, e.g., the approach in Adiwardana et al., 2020). Instead of needing a separate hand-crafted module for natural language understanding, dialogue management, etc., modern chatbots seek to model the flow of conversation and wrap up the components into a single end-to-end model (Henderson et al., 2018). This type of machine-learned model allows open-domain systems to be adaptive and respond in interesting and surprising ways.

On the other hand, task-oriented systems are limited in scope, and painstaking to create because of their ontology-based approach (Lubis et al., 2020). Yet, they are more commonly used in commercial applications both because their function is to help users and because their strict ontologies ensure safe and effective output. The hand-crafted ontologies ensure that the system will not misrepresent the company by producing nonsensical or ungrammatical responses, using offensive language, giving inaccurate information, etc. Further, customers are mainly looking for solutions and results in customer service applications, not chitchat.

However, both systems fail to escape the "uncanny valley", a significant challenge of conversational systems (Lubis et al., 2020). The uncanny valley describes the disconcerting space when a system is approaching human-likeness but falls just short enough to make the user lose trust in the system and feel uncertain. It lies between systems that truly act human-like and those that are overly simplistic and fake. For example, some chatbots attempt to avoid this condition by being strategically vague, thereby eliminating the possibility of a contradiction (Adiwardana et al., 2020). At the other end of the spectrum are the virtual assistants that have hard-coded responses to a variety of inputs but can't progress a conversation any further.

2.2 Shortcomings in Conversational A.I.

The inability for conversational systems to communicate well enough to be perceived as human has many contributing factors. This section will attempt to cover some of the more salient shortcomings in the field, beginning with the most egregious, the machine learning approach, and moving on to discuss concerns with data and evaluation.

2.2.1 Problem of Machine Learning Approaches

At the most fundamental level, conversational systems—whether task-oriented or otherwise—fail to achieve human parity because we don't understand conversation either. In his longitudinal review of computational linguistics, Church (2011) details the tension between the empiricist and rationalist perspectives. Broadly speaking, the rationalist approach deals with formal linguistic structures while the empiricist approach uses machine-learning algorithms to make a best guess. In his definition, the current data-driven machine learning approaches to NLP (natural language processing) would be considered empiricist as they use empirical methods to approximate natural phenomena. The rationalist perspective attempts to create models to replicate the true nature of things. For example, in the related field of speech synthesis, a rationalist would attempt to create a computational model of the human articulatory system. On the other hand, an empiricist would take the pragmatic approach of generalizing from existing data e.g., concatenating phonemic recordings (Church, 2011). This pragmatic approach often leads to successful results quickly, even if it does not address the underlying systems at play.

Likewise, the field of NLP shifted from a rationalist approach with the understanding that empirical methods are not perfect but give accessible and acceptable results until our understanding grows (Church, 2011). The problem is that in recent years, researchers have forgotten this fact and have focused primarily on adding more data and more power to the best models of the day (see, e.g., the approach in Adiwardana et al., 2020).

Back in the 1960s, it was shown that the perceptron, and by extension most modern machine learning algorithms, cannot learn functions that are not linearly separable (Minsky and Papert, 2017; for further discussion, see Church, 2011). For example, the natural language understanding tasks of coreference resolution and word sense disambiguation are not linearly separable. Continuing this argument, Saba (2021) argues that not only is natural language understanding a challenging task for machine learning but that it is fundamentally contradictory. At a high level, he argues that humans compress their speech to omit common knowledge elements for efficiency. Thus, much of the real meaning in an utterance never appears in a natural language corpus. As a result, ML algorithms are not trained on, and therefore cannot learn to understand, the actual meaning of a given utterance.

Similar concerns are raised about the ability of a machine learning system to adapt to new scenarios. Ben-David et al. (2019) showed mathematically that an algorithm can only learn if the data is sufficiently compressible, that is, if it's highly redundant. This need for redundancy gives rise to the old tongue-in-cheek adage, "There's no data like more data". However, in practice, the amount of redundancy required is somewhat extreme; to achieve human-like accuracy in some tasks, super-human amounts of data are required. For example, the current state-of-the-art speech recognition systems, like the recent HuBERT model (Hsu et al., 2021), can achieve a word error rate of less than 2% by leveraging tens of thousands of hours of speech data, millions of words of text and massive amounts of computational might. According to Moore's (2003) estimations, the average person won't hear this much speech until they are in their thirties!

It should be noted that there is merit to the data-driven approach. In a recent paper, the Google Brain Team (Adiwardana et al., 2020) tested the hypothesis that a better conversational system could be achieved merely by adding more data and features to a computational model. They created Meena, a chatbot trained on 867 million context response pairs using a 2.6 billion parameter sequence to sequence model. Using their custom metric, SSA (discussed in Section 2.2.3), Meena scored 79% over seven conversational turns, while humans scored 86%. This score was over 20% better than every other open-source chatbot they assessed, supporting the data first hypothesis. However, rationalist Pierce (1969) would call this an "artful deceit" that tricks users into believing the system understands more than it actually does.

2.2.2 Data Issues

Moving on from the fallacies of machine learning as an approach, we note that the amount of data needed for state-of-the-art speech recognition does not even exist for conversations. Finding quality datasets that are labeled with relevant features is a significant problem in the field (Dai et al., 2020). This problem compounds when a specific task is modeled on top of the inherent conversational structure modeling needed for a system.

As a result, each research study uses a different data source, often creating custom datasets, to generate their results.

Some examples of common data sources and studies that use them are:

- social media (Neff and Nagy, 2016; Zheng et al., 2020; Adiwardana et al., 2020)
- book corpora (Wolf et al., 2019)
- movie corpora (Liu et al., 2017)
- small recorded conversation competitions (Fulda et al., 2018)
- crowd-sourced methods (Kelley, 1984; Wen et al., 2017; Budzianowski et al., 2018)

These data sources can be useful in the correct context but come with unique shortcomings. For example, social media datasets have the benefit of being easily obtainable, making them some of the largest datasets. However, there are differences between how people speak online and in person, limiting this data's efficacy (Bowden et al., 2019). Even more critically, social media is permeated with highly controversial material, personal insults, incendiary opinions, and unnecessarily vulgar utterances (Neff and Nagy, 2016; Fulda et al., 2018).

Similarly, book and movie corpora are also easily accessible and tend to be cleaner in terms of content and the amount of preprocessing needed. Yet, linguistic patterns in book and other general language corpora are very different from those present in conversations making it a poor model for conversation (Bao et al., 2020). Movie dialogue tends to be overly dramatic and focus on onscreen elements making conversational utterances incomprehensible out of context (Fulda et al., 2018). Competitions and crowd-sourced methods have the blessing and curse of being created in a structured environment. As a result, corpus builders can be intentional about the type of content present, e.g., task-oriented, non-offensive, etc. However, as a result, they lack the naturalness of spontaneous speech and are limited in scope.

At a higher level, the existence of different types of datasets is positive as it enables concurrent innovation in multiple areas. The concern is that each study listed by a data type above needed to create their own data. It is impossible to facilitate collaboration between researchers without a way to compare results. This trend is beginning to change as more studies publish their datasets (see, e.g., Lowe et al., 2015; Wu et al., 2020).

2.2.3 Concerns about Evaluation

Further, once useful data is acquired, there is no straightforward metric to rate a conversation. In general, users expect a conversational system to be human-like. However, human-likeness is a broad and abstract ideal that is often subjectively defined (Adiwardana et al., 2020). Additionally, different types of metrics are needed for different tasks. Chatbots try to be human-like, while in a purely task-oriented system, the most relevant metric might be task success or user satisfaction (Jurafsky and Martin, 2021). In this section, I will highlight common evaluation metrics and their shortcomings and then detail a recent metric that seems promising.

Historically, chatbots have been trained to pass the Turing Test—originally called the Imitation Game (Turing, 1950). In this test, a judge questions a subject for five minutes and then guesses whether or not the subject is human. However, this is a subjective measurement and costly as it is determined by humans (Lowe et al., 2017). Even more crucially, the binary determination of the Turing Test is not a metric that can be used to compare similar models. The BLEU score for machine translation (Papineni et al., 2002) is commonly used as an automatically generated (as opposed to subjective human-generated) metrics in conversational systems (Lubis et al., 2020). However, this metric compares the similarity of generated text with the data. Hence, it merely serves as an abstract gauge of human-likeness that doesn't correlate with human's perceptions of conversational quality (Liu et al., 2016).

Similarly, word embedding type approaches similar to Word2vec (Mikolov et al., 2013) are commonly used to compare generated utterances with those found in human speech. However, Ghandeharioun et al. (2019) showed that this type of metric also doesn't correlate well with human evaluations. Nonetheless, BLEU and embedding space similarity remain the most common metrics used because there just aren't better options (Lowe et al., 2017).

As mentioned earlier, in the same paper where they proposed Meena (2020), Adiwardana et al. also proposed the Sensibleness and Specificity Average (SSA). This is another human-generated metric where each utterance is given a binary score whether or not it was Sensible in the context and separately a binary Specificity score to measure against vague replies. After each conversation, cumulative scores for both are Averaged together. As shown in Figure 2, Adiwardana et al. (2020) found this score correlated highly with human-likeness. This correlation is likely because it grades the relevance of responses in context as well as docking responses that were too generic, e.g., "I don't know."

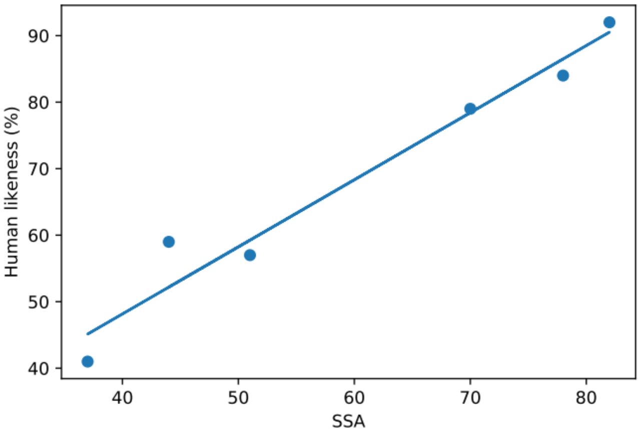

Figure 2: A plot of the SSA against human-likeness (with a regression line to show correlation) as scored by crowd-sourced judges in (Adiwardana et al., 2020)

Interestingly, chatbots trained to dupe the Turing test don't consistently score well on this metric as their success lies in strategic vagueness (Adiwardana et al., 2020). While it is a human-generated metric—and not automatic like the ones above—I believe that the SSA is the current best metric for evaluating conversational systems of any type. Encouragingly, even though it was proposed recently, it is already starting to gain traction in new work (e.g., Gerhard-Young et al., 2021).

Similar to the dataset problem, this variety of metrics means that there is no way to accurately compare multiple systems' success. In other domains, models are compared by their accuracy on a given dataset (e.g., Ruder, 2021). This ability gives other fields a measure of cohesiveness and allows researchers to build off each other's work. However, as discussed, consistent corpora and metrics are only starting to be introduced to the field of conversational A.I. As a result, researchers essentially have to start from scratch in each project and have no way of validating the veracity of other results.

2.3 Ethical Concerns

Notwithstanding the aforementioned problems with conversational A.I., both as a field and in practice, conversational agents are in widespread use today. Though they are not truly conversational machines, virtual assistants like Siri, Alexa, and the Google Assistant reach millions of users. According to one study, there are over 110 million virtual assistant users in the United States alone (Liu, 2020). However, there is danger in treating these assistants as reputable sources of knowledge. For example, ELIZA (Weizenbaum, 1966) took on the persona of a psychotherapist, and many participants had trouble believing that it was not human. In a more recent study, Bickmore et al. (2018) tested the implications if people were to treat their virtual assistants as sources of valid medical advice. Participants asked virtual assistance for medical advice arising from an everyday situation. Bickmore et al. (2018) found that in the sensible responses, almost a third of the suggestions would cause the patient harm, and many could be fatal.

In addition to harmful advice from conversational applications, chatbots are also prone to generating harmful and offensive utterances. Perhaps the most infamous example is the 2016 Microsoft chatbot, Tay (Lee, 2016). This conversational system was released publicly on Twitter and then removed after sixteen hours for obscene and inflammatory tweets (Neff and Nagy, 2016). This bot was trained on social media data which, as mentioned in 2.2.2, is a popular source for generating dialogue datasets. However, as the Microsoft researchers discovered, social media is rife with toxic and abusive language, which the chatbots then mimic in unfortunate ways. (Henderson et al., 2018). This is likely why (Adiwardana et al., 2020) did not publicly release Meena to the general public and why this type of end-to-end bot is less common in industry (Nancy Fulda, p.c., December 29, 2021).

Another significant concern with conversational systems is their ability to maintain privacy (Henderson et al., 2018). This was seen from the early days of ELIZA (Weizenbaum, 1966), where it was noted that people would freely share intimate details of their lives with the machine (Weizenbaum, 1976). All user inputs are recorded of necessity for use in the natural language understanding unit of a conversational system. From there, it is up to the system designer to determine if, where, and how this input data is stored. Thus, privacy is of paramount concern today with the smart devices in people's homes always listening for their command words.

2.4 Recap on the background

Success in other branches of artificial intelligence has been the result of improved learning algorithms and publicly distributed datasets (Lowe et al., 2015). However, as we have seen, machine learning approaches to conversational A.I. are not the golden solution they are often touted to be. These large models have their place, but it seems clear that we will not achieve any measure of human parity with empiricist approaches alone. For conversational A.I. to move forward, we need a system that can reintroduce linguistic structure while leveraging the adaptability that ML promises.

Further, to avoid the uncanny valley, we need a system that can reliably stay on task and help users quickly while still having the flexibility to add naturalness and flavor to the conversation. The current lack of unity in conversational A.I. research inhibits these lofty goals. We need to start doing benchmark testing on standard datasets with standard metrics.

Lastly, only a very few of the research papers cited in this work acknowledge the ethical concerns of conversational A.I. If the field is to move forward in a safe and user-friendly way, researchers need to address and respond to the ethical concerns inherent in their work.

3. Methods

As noted in Section 2.2.1, the pragmatic empiricist approach produces better results in the short term (Church, 2011) and tends to advance more quickly than true science (Pierce, 1969). As a result, none of the research in this section could be considered as coming from a rationalist perspective. However, I believe that the methods presented below are more promising than a pure machine learning approach because, like humans, they combine multiple elements into one system.

The methods presented fall into two major categories, modular and pretrained approaches. In general, any system based on the framework illustrated in Figure 1 could be considered modular. Those presented here differ in that the source of their generated output is modular, enabling greater scalability. The pretrained approaches are far more empiricist but show promising work into applying the recent field-altering successes of pretrained models like BERT (Devlin et al., 2019) and the GPT models (Radford et al., 2018, 2019; Brown et al., 2020).

3.1 Modular Approaches

3.1.1 A Network-based End-to-End Trainable Task-oriented Dialogue System

Noticing similar gaps in the field that I called attention to in Section 2.2, Wen et al. (2017) created a modular ML framework for task-oriented dialogue. Their work is particularly exciting because, unlike almost all task-oriented systems, they essentially eliminated the costly ontology generation step. They did this by creating a training dataset with learnable features. Further, this dataset was relatively cheap and clean because they combined the Wizard of Oz (WOZ) paradigm (Kelley, 1984) with Mechanical Turk to crowd-source a conversational dataset crafted toward their specific task. Thus, they could use data generated models for the natural language understanding, belief tracking, and natural language generation still including slots for the to model point to a database to provide relevant, accurate information. Their completed dataset contained approximately 680 dialogues, totaling 1500 individual dialogue turns.

Like the SSA, Wen et al. (2017) used crowd-sourced evaluators who were asked to evaluate the model's task success rate (a binary score) and give scores out of five for comprehension ability and naturalness. Even with the relatively small size of their corpus, they found that it could complete 98% of tasks successfully and comprehension ability and naturalness scores averaged more than 4 out of 5.

Wen et al. (2017) continue the unfortunate trend of generating their own dataset and evaluation metrics. However, they have made both publicly available for other researchers to continue their work (as done in Wu et al., 2020). Further, while they use machine learning methods in almost every model component, they used a novel method of connecting the learned features with real word databases to ensure accurate output. As a result, their task-oriented system is more adaptable and presumably more scalable than most.

3.1.2 BYU-EVE

Similar to Wen et al. (2017), Fulda et al.'s (2018) BYU-based team realized that a single ML approach would be insufficient to approximate human conversation. However, they took the modularity a step further by creating a system of response generators instead of the one model connected to a database. For each user input, their conversational model, EVE, attempts to generate an emotive and a factual response as well as an utterance that helps the conversation move forward. Each of these responses can be generated by one of several distinct models, an API to knowledge sources, and the knowledge graph created during the conversation. Each proposed response is scored or merged with other responses to generate the final output.

Another interesting aspect of their work is that where other groups might use a sequence to sequence model to generate responses and score outputs, they used conversational scaffolding (Fulda et al., 2018). Conversational scaffolding is a direct response to the seeming inability of neural approaches to filter out offensive utterances. This approach uses a sentence-level embedding space (as opposed to the word-level embeddings in Word2vec (Mikolov et al., 2013)) to rank generated responses based on how similar they were to something a human might say. This approach assumes that there is no single correct response to an utterance, so the system needs only to find one with suitable characteristics.

Knowing that generated responses would end up in a similar style to the training set, the researchers needed a high-quality dataset. Unlike most open-domain system developers, the BYU team was loath to use a conversational corpus from social media data because of the high frequency of offensive content. Instead, they created a conversational competition where participants who agreed to abide by a high level of conduct were paired with conversational partners. Rating participants based on their comments' length and style encouraged engaging and content-rich utterances. Their dataset contains over 90,000 utterances from almost 1200 different users.

EVE (Fulda et al., 2018) scored admirably in its evaluations. It scored an average of 3.25 out of 5 on the subjective customer satisfaction rating. Though this is not the most impressive result, when the response wasn't optimal, EVE "failed with dignity" (Fulda et al., 2018), meaning that its responses were usually still satisfactory.

More than any other model cited in this review, EVE (Fulda et al., 2018) has the impressive ability to leverage strengths from a variety of areas. It is an open-domain system but done in such a way that it could presumably be adapted to be task-oriented as well. It is data-driven without generating ethically concerning responses. Further, it can combine different styles of responses to respond in a human-like way. Lastly, even though they too continued the trend of using a custom dataset, Fulda et al. (2018) also released it publicly so that others could continue their efforts.

3.2 Task-Oriented Pretrained Models

Unlike the modular approaches presented in Section 3.1, this section focuses on end-to-end approaches to task-oriented dialogue models. It is well known in the field of NLP that using large pretrained models like GPT (Radford et al., 2018, 2019; Brown et al., 2020) and BERT (Devlin et al., 2019) tends to improve the accuracy of most NLP tasks. While this type of approach directly contradicts the rationalist perspective (as discussed by Church, 2011; Saba, 2021), it follows the current trends of the field and so is worth discussing here. The models presented in this section propose initial work to apply this model to task-oriented dialogue problems.

3.2.1 Hello, It's GPT-2

In their paper, "Hello, It's GPT-2 - How Can I Help You?", Budzianowski and Vulić (2019) provide an initial exploration into finetuning GPT-2 (Radford et al., 2019) for task-oriented dialogue. Their work builds on previous work by Wolf et al. (2019), which uses the GPT (Radford et al., 2018) architecture to create an open-domain chatbot. Budzianowski and Vulić (2019) use a sequence to sequence model to adapt the task-oriented dialogue framework to operate entirely on text input. As a result, they avoid creating an explicit dialogue management module and a domain-specific natural language generation module, which greatly simplifies the process of creating a task-oriented dialogue system.

To ensure that the system produces domain-specific responses, they include the belief state and knowledge base representation along with the user's utterance as input (Budzianowski and Vulić, 2019). They found that this method makes their model highly portable to different domains while still generating relevant responses.

While the results are preliminary, they show a matched or improved score (both automatically and human-generated) on most tasks.

3.2.2 TOD-BERT

Wu et al. (2020) build on the results of Budzianowski and Vulić (2019) by building their own pretrained model instead of just finetuning another model. They note that most of the recent groundbreaking results in NLP have come from finetuning the seminal pretrained models mentioned. Thus, instead of finetuning those general language models, they create a new model built on task-oriented dialogues. As a result, others will have a domain-specific model to finetune in the future. They call their model TOD-BERT because it is a new version of BERT (Devlin et al., 2019) trained on Task-Oriented Dialogues (TOD).

Wu et al. (2020) seek to prove that a self-supervised language model pre-training using task-oriented corpora can perform better in task-oriented tasks than existing pre-trained models. Thus, they follow the BERT's (Devlin et al., 2019) pre-trained model architecture as closely as possible to provide an accurate comparison. However, unlike BERT-which was trained primarily on Wikipedia data (Devlin et al., 2019)-Wu et al. (2020) use only open-source human-human task-oriented datasets.

Their results surpassed all other systems in the central areas of natural language understanding, dialogue tracking and prediction, and response selection. Further, as they intended to build a pretrained system to support further work, their model and all associated work are available online.

4. Proposals

These prior works demonstrate exciting new approaches to improving the field of conversational A.I. While there are other notable developments in the field, each of the works presented in Section 3 focuses on areas or approaches that I believe are particularly valuable. In this section, I propose ideas for future studies that would build on the successes of these works.

4.1 Response Evaluator Research

While each of the above models offers a new approach to building and simplifying task-oriented systems, none of them go back to the rationalist approach argued for by Saba (2021). Unfortunately, Saba did not give many suggestions for what this approach could entail. Further, it is unclear whether there is enough knowledge about how conversation works to shift from the current empirical approaches. It is clear, though, from the discussion about evaluation methods in Section 2.2.3 and my own exploration, that a response evaluator is needed. If a model could rank prospective responses with absolute surety, then conversational A.I. would be ready for the self-play reinforcement learning technique that was so successful with AlphaGo (Silver et al., 2016).

Answering the question of creating the perfect response evaluator for single utterances in context, and by extension, evaluating the success of a computational system, is too big a place to start. Presumably, we would first need to understand how humans respond to speech. However, one interesting approach to begin exploring this problem would be to create some computational ranking system based on the Gricean maxims (Grice, 1975). In this seminal work, Grice lays out several rules that are inherently understood in conversations, and teaching a computer to obey the maxims would definitely move the field forward.

Some prior work has focused on creating an automatic metric from the Gricean maxims, but it has not received mainstream attention. For example, Jacquet et al. (2019) studied the impact of violations of these maxims on the humanness of a system, but did not propose a metric. Additionally, other studies have shown the efficacy of the maxims for evaluating chatbots (Saygin and Cicekli, 2002; Chakrabarti and Luger, 2012; Jwalapuram, 2017; Ngai et al., 2021) but primarily focused on qualitative approaches. I propose a research study that builds on these results by training a model to score utterances on each of the four aspects of Grice's maxims (1975). Like the former studies, this model would be trained on crowd-sourced judgments for both human and machine-generated dialogues. This Gricean evaluation model would then be used to compare the generated utterances of a variety of open access chatbots.

A study of this sort would progress the field of conversational A.I. forward in significant ways. First, it would propose a baseline metric for evaluating conversations. Additionally, it has the benefits of solid linguistic backing and being automatically computed (as opposed to human judged). Lastly, this study would show qualitatively how current conversations compare with each other.

4.2 A New Model

Like Saba (2021), I don't believe that a single machine learning approach is sufficient to solve the problem of natural language. It seems that instead of building bigger and bigger neural end-to-end models, we should focus our efforts on approaches that combine multiple approaches. Accordingly, I propose the creation of a new task-oriented system drawing on the best methods from the abovementioned papers. This novel model will be distinguished from other approaches by allowing multiple generator modules like the BYU-EVE system (Fulda et al., 2018). Unlike EVE, however, these generators will respond to both task-oriented and conversational intents. I will attempt to connect multiple open-source chatbot models like TOD-BERT (Wu et al., 2020), Pender Bot (pender, 2018), and API Generators to sources like ESPN, Washington Post DuckDuckGo, Wikipedia, etc. Each of these sources has different abilities and uses. TOD-BERT (Wu et al., 2020) represents a great effort to harness the power of BERT for task-oriented dialogue, while Pender Bot (pender, 2018) is a fun but highly rated open domain bot. The API Generators will allow it to share real-world information from various sources.

Since these sources already exist, my primary contribution will be to create a response evaluator to choose the best utterance between the different generators. For the Natural Language Understanding unit, I will follow Wen et al.'s (2017) neural intent classification approach. They used an LSTM (Long Short-Term Memory) model to encode utterances into features from which they determined the correct intent. From here, the intent will be passed to each generator to propose responses. Choosing which response to use is the tricky part. Fulda et al. (2018) used a sentence embedding approach to choose the most human-like and relevant utterance. Others have used seq2seq models here. I don't believe that either method alone is sufficient, so I propose to use a combination of the two approaches.

The details of this combination effort remain unclear, but I will attempt to build on prior studies. For example, this combination effort has been explored previously in the Eighth Dialogue System Technology Challenge (Gu et al., 2021). For the challenge, Gu et al. explored combining token-level word embeddings with a seq2seq model that they discussed, and I could attempt to apply it here. Other possible sources to consider are Yan et al. (2017) or Whang et al. (2021). Alternatively, I could use a response evaluator like the one considered in Section 4.1.

Lastly, the quality of the resulting system should be evaluated using the SSA metric (Adiwardana et al., 2020). This choice is both because I believe this is the best metric currently in use and to help it to gain traction. Together, the ideas presented in this section should add a new perspective to the inspiring works presented in Section 3 and continue the much needed research into response evaluators.

5. Conclusion

The field of conversational A.I. has a long and rich history but still fails to reach the level of coherency seen in other fields. However, even with its current shortcomings, conversational systems remain an exciting field of research with promising areas of new development. While not precisely what Saba (2021) and other rationalists envisioned, the modular approaches proposed by Fulda et al. (2018) and Wen et al. (2017) enable modern conversational systems to be adaptable and give more nuanced output. While the sequence to sequence models proposed by Wu et al. (2020) and Budzianowski and Vulić (2019) lean into the ML approach, they do it in a more promising way than the pure end-to-end systems. Drawing upon the strengths of these approaches and adding innovations of their own, future researchers will yet achieve Turing's dream of truly conversational machines.

5.1 Future Work

Other than the proposals, there is still work to be done in this field. For example, the current work assumed a text-based conversational environment with clean input. Adding the possibility of errors in the text input, especially those caused by speech recognition systems, introduces errors that can be propagated through the pipeline (Lubis et al., 2020). Future work would address methods to respond to this and similar drawbacks, making speech-based conversational systems more robust.

5.2 Acknowledgements

Many people have been essential to my learning about the ideas presented in this work. I would like to specifically thank Seth Levine for encouraging me to think about the shortcomings of conversational systems and introducing me to Saba (2021). Additionally, my conversation with Marie Meteer forced me to think critically about what form a hybrid rule-based and neural model might take. Lastly, I would like to thank Nancy Fulda for meeting with me to help me understand some of the more complicated elements of EVE (Fulda et al., 2018) and focusing my thoughts on important next steps for the field.

Works Cited

Daniel Adiwardana, Minh-Thang Luong, David So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc Le. 2020. *Towards a Human-like Open-Domain Chatbot*. January.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online, July. Association for Computational Linguistics.

Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. 2019. Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48, January.

Timothy W. Bickmore, Ha Trinh, Stefan Olafsson, Teresa K. O'Leary, Reza Asadi, Nathaniel M. Rickles, and Ricardo Cruz. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research*, 20(9):e11510, September.

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. GUS, a frame-driven dialog system. *Artificial Intelligence*, 8(2):155–173, April.

Kevin K. Bowden, Shereen Oraby, Amita Misra, Jiaqi Wu, Stephanie Lukin, and Marilyn Walker. 2019. Data-Driven Dialogue Systems for Social Agents. In Maxine Eskenazi, Laurence Devillers, and Joseph Mariani, editors, *Advanced Social Interaction with Agents*, volume 510, pages 53–56. Springer International Publishing, Cham.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong, November. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October. Association for Computational Linguistics.

Chayan Chakrabarti and George F. Luger. 2012. A semantic architecture for artificial conversations. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 21–26, Kobe, Japan, November. IEEE.

Kenneth Church. 2011. A Pendulum Swung Too Far. *Linguistic Issues in Language Technology*, 6(5):1–27, October.

Yinpei Dai, Huihua Yu, Yixuan Jiang, Chengguang Tang, Yongbin Li, and Jian Sun. 2020. A Survey on Dialog Management: Recent Advances and Challenges. , May.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Nancy Fulda, Tyler Etchart, William Myers, Daniel Ricks, Zachary Brown, Joseph Szendre, Ben Murdoch, Andrew Carr, and David Wingate. 2018. BYU-EVE: Mixed Initiative Dialog via Structured Knowledge. In *Proceedings of the 2018 Amazon Alexa Prize*. November.

Greyson Gerhard-Young, Raviteja Anantha, Srinivas Chappidi, and Björn Hoffmeister. 2021. Low-Resource Adaptation of Open-Domain Generative Chatbots. *arXiv preprint arXiv:2108.06329*, August.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. In *Advances in Neural Information Processing Systems*, volume 32, Red Hook, NY, USA, December. Curran Associates, Inc.

H. P. Grice. 1975. Logic and Conversation. *Speech Acts*:41–58, December.

Jia-Chen Gu, Tianda Li, Zhen-Hua Ling, Quan Liu, Zhiming Su, Yu-Ping Ruan, and Xiaodan Zhu. 2021. Deep Contextualized Utterance Representations for Response Selection and Dialogue Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2443–2455.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, New Orleans LA USA, December. ACM.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech RepresentationLearning By Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1, October.

Baptiste Jacquet, Alexandre Hullin, Jean Baratgin, and Frank Jamet. 2019. The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots. In *2019 International Conference on Information and Digital Technologies (IDT)*, pages 180–189. June.

Dan Jurafsky and James H. Martin. 2021. *Speech and Language Processing*. 3rd ed. draft edition, September.

Prathyusha Jwalapuram. 2017. Evaluating Dialogs based on Grice's Maxims. In *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 17–24, Varna, September. INCOMA Ltd.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1):26–41, January.

Peter Lee. 2016. Learning from Tay's introduction. March.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November. Association for Computational Linguistics.

Huiting Liu, Tao Lin, Hanfei Sun, Weijian Lin, Chih-Wei Chang, Teng Zhong, and Alexander Rudnicky. 2017. RubyStar: A Non-Task-Oriented Mixture Model Dialog System. In *Proceedings of the 2017 Amazon Alexa Prize*, page 10. arXiv preprint, November.

Shanhong Liu. 2020. U.S. voice assistant users 2017-2022. December.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada, July. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September. Association for Computational Linguistics.

Nurul Lubis, Michael Heck, Carel van Niekerk, and Milica Gašić. 2020. Adaptable Conversational Machines. *AI Magazine*, 41(3):28–44.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013, January.

Marvin Minsky and Seymour A. Papert. 2017. *Perceptrons, Reissue of the 1988 Expanded Edition with a new foreword by Léon Bottou: An Introduction to Computational Geometry*. MIT Press, September.

Roger K. Moore. 2003. A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners. In *8th European Conference on Speech Communication and Technology*, pages 2582–2584, Geneva, Switzerland.

Gina Neff and Peter Nagy. 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10:4915–4931, October.

Eric W. T. Ngai, Maggie C. M. Lee, Mei Luo, Patrick S. L. Chan, and Tenglu Liang. 2021. An intelligent knowledge-based chatbot for customer service. *Electronic Commerce Research and Applications*, 50:101098, November.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

pender. 2018. *chatbot-rnn*. February.

J. R. Pierce. 1969. Whither Speech Recognition? *The Journal of the Acoustical Society of America*, 46(4B):1049–1051, October.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI Blog*:12.

Alec Radford, Jeff Wu, R. Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.

Owen Rambow and Tanya Korelsky. 1992. Applied text generation. In *Proceedings of the third conference on Applied natural language processing*, pages 40–47, USA, March. Association for Computational Linguistics.

Sebastian Ruder. 2021. Tracking Progress in Natural Language Processing.

Walid Saba. 2021. Machine Learning Won't Solve Natural Language Understanding. *The Gradient*, August.

Ayse Pinar Saygin and Ilyas Cicekli. 2002. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258, March.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January.

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

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

Joseph Weizenbaum. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman, San Francisco.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for*

Computational Linguistics: Volume 1, Long Papers, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do Response Selection Models Really Know What's Next? Utterance Manipulation Strategies for Multi-turn Response Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14041–14049, May.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *arXiv preprint arXiv:1901.08149*, January.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online, November. Association for Computational Linguistics.

Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–694, New York, NY, USA, August. Association for Computing Machinery.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027, October.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2020. Personalized Dialogue Generation with Diversified Traits. *arXiv:1901.09672 [cs]*, January.