Discourse over Discourse: The Need for an Expanded Pragmatic Focus in Conversational AI

Anonymous ACL submission

Abstract

The summarization of conversation, that is, discourse over discourse, elevates pragmatic considerations as a pervasive limitation of both summarization and other applications of contemporary conversational AI. Building on impressive progress in both semantics and syntax, pragmatics concerns meaning in the practical sense. In this paper, we discuss several challenges in both summarization of conversations and other conversational AI applications, drawing on relevant theoretical work. We illustrate the importance of pragmatics with so-called star sentences, syntactically acceptable propositions that are pragmatically inappropriate in conversation or its summary. Because the baseline for quality of AI is indistinguishability from human behavior, we draw heavily on the psycho-linguistics literature, and label our complaints as "Turing Test Triggers" (TTTs). We discuss implications for the design and evaluation of conversation summarization methods and conversational AI applications like voice assistants and chatbots.

1 Introduction

The summarization of conversation, a case of discourse over a discourse, clearly illustrates a series of pragmatic limitations in contemporary conversational AI applications. While there has been some previous work examining pragmatic issues in conversational AI (i.e., (Bao et al., 2022; Kim et al., 2020, 2021a; Nath, 2020; Wu and Ong, 2021)), additional progress depends on understanding the source of limitations in current applications. We aim to contribute to both theory and applications by examining several recurrent pragmatic limitations associated with both conversation summarization models and other conversational AI applications.

No doubt, applications of conversational NLP have achieved considerable performance improvements and are an area of increasing research focus. Deep learning has enabled the development of personal assistants (Radford et al., 2019), ASR (Hinton et al., 2012) and machine translation (Sutskever et al., 2014), as well as large language models that can generate seemingly natural sentences (Brown et al., 2020; Devlin et al., 2019; Zhang et al., 2022b). This work has spurred the development of models that can employ certain characteristically human aspects of dialogue (Kim et al., 2020; Majumder et al., 2020; Wu et al., 2019). Likewise, automatic summarization has expanded from sentence compression and document summarization (Bhandari et al., 2020; Li et al., 2017; Nayeem et al., 2018; Wang et al., 2020) to summaries of casual conversations (Chen and Yang, 2021; Goo and Chen, 2018) and meetings (Gillick et al., 2009). Such summaries are intended to provide an account of 'what the exchange was about'. Moreover, performance of these models has improved, prompting empirical work on new evaluation metrics (Bhandari et al., 2020). Despite these performance gains, there are remaining areas for improvement in both conversational summarization and conversational AI more broadly. We illustrate the remaining challenges in this area with ill-conceived examples inspired by conversational AI systems (Gratch et al., 2014), conversation summarization models, (Gaur et al., 2021) and author interactions with chatbots and voice assistants. Like Chomsky's star sentences, these examples have clear pragmatic deficiencies that trigger the Turing Test criterion. No competent speaker would construct such discourse.

Our approach helps structure user frustrations with conversational AI systems documented in the HCI literature, typically highlighting complaints about conversational skills separately from other usability concerns (Brandtzaeg and Følstad, 2017; Følstad and Skjuve, 2019; Liao et al., 2016; Luger and Sellen, 2016; Porcheron et al., 2018; Zamora, 2017). In so doing, we promote synergy between applied and basic research endeavors that address language in use. In particular, we suggest below that pragmatic limitations (and user frustrations) of current conversational AI systems are addressed by the pragmatic theory of *relevance*. Next, we articulate two sub-themes for understanding and addressing these limitations: preservation of meaning and incorporation of external context. Throughout, we inventory a series of pragmatic failures in different domains, and discuss them in the context of relevant theoretical work in linguistics, psycholinguistics, and cognitive psychology.

2 User Relevance

Conversations (and comprehensive summaries) preserve *relevance*. When people engage in conversation, they expect that their conversational partners will make contributions that are relevant to the ongoing dialogue and consistent with the accepted purpose of the conversation. When users interact with conversational AI applications, they have similar expectations for those systems (i.e., (Zamora, 2017)). (Grice, 1975) initially proposed that the expectation of relevance is due to a *cooperative principle* and that the expectation of related utterances is due to a maxim of *relation*.

(Sperber and Wilson, 1986) revised this explanation. Instead of an explicit cooperative principle, they proposed that the search for relevance is a basic feature of human cognition (Wilson and Sperber, 2013). A given input is relevant if processing that input generates a worthwhile difference to a recipient's representation of the world (Wilson and Sperber, 2013). Relevance is necessarily determined in context. For a given stimulus to be relevant in the prevailing context, it must be worth the recipient's processing effort and be the most relevant stimulus available that is compatible with the person's abilities and preferences (Wilson and Sperber, 2013). Given that all other factors are equal, if one stimulus requires less processing effort than another, that stimulus will be the most relevant.

Relevance theory emphasizes that any content a conversational AI system provides that is difficult to read, missing important information, or is incorrect reduces relevance to the end user. When users encounter this type of content, they must either expend additional effort to understand the information or search elsewhere for more appropriate information. Consider a source conversation and its attempted automated *star* summary below.

Interviewer: Can you give me an example of that?

Patient: If somebody ... annoys me I'll probably let them know they're annoying me until they stop.

(TTT) Patient: Yeah

Pragmatically-appropriate summary: I would let them know they're annoying me.

The hypothetical summary of the participant response, *yeah*, is not be acceptable either in the source conversation or as a summary. From a theoretical perspective, such a response is not pragmatically appropriate. The interviewer's question is an example of an indirect speech act (Clark, 1979) with both literal and indirect meaning.

An answer that addresses just the literal form of the question, as in the example, violates the expectation of relevance. The intended, indirect meaning of the question requires a response that addresses that meaning. The direct meaning response, either in actual conversation or it's summary, creates the impression of a flippant patient who flouts relevance expectations. Indeed, (Clark, 1979) demonstrated that responses to indirect speech acts predominantly answer the indirect meaning, as in the above actual response (Clark, 1979). From a practical perspective, the suggested summary fails to convey the patient's actual response.

A summarization method that does not preserve the original meaning of the conversation risks a false conclusion (Wilson and Sperber, 2013). Should a user recognize that this answer is likely misleading (because it violates expected responses of indirect speech acts), the user must revisit the original conversation to determine the correct account. This failure of relevance owing to the associated additional effort fatally reduces the utility of the summarization model. The example above, and related empirical work in HCI, illustrates two fundamental and related dimensions of relevance; meaning and contextual awareness. Next, we examine these two dimensions more closely and pragmatic failures that can arise from each.

3 Meaning and Inference

In order to generate relevant content, conversational AI systems must respond to the meaning of a user's utterance, with respect to its contribution to the conversation as a whole. In this section, we discuss some pragmatic aspects of meaning as they relate to conversational AI.

Meaning is often reduced to the domain of classical semantics in philosophy. Accordingly, discourse meaning lies in propositional content. *Propositions* are the purported elementary units of meaning and represent truth measurable in the world (Levinson, 2011). The flawed summary above omitted its explicit propositional content. From a psychological perspective, propositions bridge the distinction between specific words and their corresponding concepts (Buschke and Schaier, 1979; Forster, 1970; Goetz et al., 1981; Graesser et al., 1980; Kintsch, 1974; Kintsch and Keenan, 1973).

One of Chomksy's most important insights is that sentences often contain multiple propositions with complex interrelationships and dependencies(i.e., (Forster, 1970; Graesser et al., 1980)). The multiple sentences of discourse exacerbate the problem of recovering interrelationships, otherwise known as coherence. Automated responses that are unable to account for between sentence relationships will compromise coherence (McNamara et al., 1996). The resulting additional effort impairs comprehension and challenges relevance. (Beck et al., 1991). The voice assistant example below illustrates this point.

User: I want to go to Cleveland, is there any construction that would slow down my trip?

(TTT) Assistant: Getting directions to Cleveland [does not provide traffic information].

Pragmatically-appropriate assistant: Traveling to Cleveland will take 3 hours. There are currently no traffic delays. [Pulls up directions]

In a typical automated reply, the assistant only addresses the first proposition in the user's utterance, wanting directions to a given city. The second proposition about traffic delays is ignored. Pragmatically acceptable responses require representation of both propositions and a knowledge-driven inference—the user is more generally concerned about travel time to Cleveland in which construction from traffic is merely an instance of the concern. From this perspective, an appropriate summary of the exchange is not 'the user wanted directions to Cleveland', but rather 'the user was concerned about potential travel delays on a road trip to Cleveland.'

4 Context

Context has two dimensions. The first concerns factors that are internal to the conversation. This dimension includes aspects of context that can be understood given the prior content of the conversation. The second deals with factors external to the specific lexical content of the conversation, typically features of its physical setting. Psychologists invoke the constructs of memory to explain context-related processing. One of these constructs concerns semantic memory, which contains general knowledge that enables the kind of inference just described. A second construct concerns episodic memory for specific events (Tulving, 1972) and includes associated aspects such as who, when and where (Nyberg et al., 1996). Semantic and episodic memory influences both conversational and external context.

4.1 Conversational Context

The specific topics, words and phrases that members of the conversation use determine a conversational context that participates in the comprehension of subsequent utterances. When humans process discourse, they expand a representation of the unfolding content. As comprehension proceeds, propositions combine into *macropropositions* that recursively combine to form the *macrostructure* (van Dijk and Kintsch, 1983). Speakers use and update these representations throughout conversation to understand their partners utterances and make relevant follow-up contributions (Clark et al., 1983; Isaacs and Clark, 1987; Lockridge and Brennan, 2002).

Furthermore, users want systems to exploit the kinds of information that are plausibly in conversational context. In fact, (Følstad and Skjuve, 2019) found that users cared more about such conversational abilities of chatbots than personalities or appearance. Users would like systems that can ask clarifying questions and remember previous interactions with that user, particularly if the interactions occur close together in time (Luger and Sellen, 2016). Users would like to understand how systems work (Liao et al., 2016; Zamora, 2017), what kinds of tasks they can do (Liao et al., 2016; Luger and Sellen, 2016; Zamora, 2017), and when systems acquire new capabilities (Luger and Sellen, 2016). The ideal collaborative system is aware of the user's status and intentions and responds accordingly (Liao et al., 2016) with personalized recommendations or help users consider multiple options (Følstad and Skjuve, 2019). Pragmatic failures arise when conversational AI systems lack the corresponding conceptual representations that a human partner would maintain during discourse. In the following subsections, we discuss some of these pragmatic failures.

4.1.1 Memory for Dialogue Topics

Speakers in conversation create expectations about what pieces of information are shared between all members of the conversation. This representation is called *common ground* (Clark, 1996). Common ground represents each individual's beliefs about what information their partner knows, based on community membership and past personal experience, including the experience of the ongoing exchange. Speakers develop models of what information their partners have processed and update them as the conversation progresses (Pickering and Garrod, 2004).Conversational AI applications generally do not create these kinds of representations (i.e., (Carpenter/Existor, 2022)) or do so in a limited fashion.

To generate pragmatically appropriate content, conversational AI applications need memory for topics that were previously addressed in the conversation, represented in *personal common ground*. This type of memory is episodic with respect to the conversation– it requires awareness of what topics were discussed, the answers that were provided, and who said what. When systems do not represent the content of prior conversation, they generate pragmatically inappropriate content, as we demonstrate next.

(TTT) Interviewer: Do you still go to therapy now?

Patient: I don't, I never went to therapy.

In using the word *still*, the automated interviewer presupposes previous therapy, without having established that fact in the conversation. A more appropriate question here would be to first ask if the patient has been to therapy and then ask if the patient still attends therapy, conditional on an affirmative response to the first question. Moreover, this illustrates that comprehensive summaries are predicated on coherent dialogue. A comprehensive summary of this exchange requires both information about if the patient has previously been to therapy and if the patient is currently in therapy. Given such an exchange, a proper summary mirrors the familiar phrasing of clinical notes: the patient is not currently in therapy and denies having been to therapy in the past.

4.1.2 Propose Relevant Next Turns

Similarly, systems must propose new turns that are consistent with prior conversational content. This is particularly apparent for applications that involve multiple turns of dialogue, but is evident for single turn interactions where a user repeatedly interacts with a system as well. The example below is from a mental health chatbot for practicing CBT skills. The chatbot asks the user to provide a thought and then proposes follow-up questions that are intended to help the user process the thought and make changes to their perspective.

Bot: Does this thought assume others are thinking negatively about you?

User: No

Bot: Does your thought place too much blame on others?

User: No it doesn't

Bot: Does it place too much blame on you?

User: Nope

(TTT) Bot: Can you rewrite the thought without any distortions?

User: It doesn't have any

(TTT) Bot: How did you do?

This chatbot does not account for the prior conversational context in generating the question *can you rewrite the thought without any distortions?*. A more appropriate implementation here might propose a different strategy for the user. This failure to account for prior conversation is further illustrated by the last question, which does not respond directly to the content of the previous utterance. A more appropriate utterance would account for the user's failure to answer the previous question and propose an alternative course of action. Moreover, an appropriate summary of this exchange would highlight that failure.

4.1.3 Correction of Prior Content

Systems need the ability to correct and update content that was previously introduced in dialogue. Information is introduced into common ground via a collaborative process between partners (Clark and Schaefer, 1987; Brennan and Clark, 1996) and partners may make changes or updates until all members are satisfied. Moreover, speakers routinely engage in self-correction to revise prior mistakes (Schegloff et al., 1977). Conversational AI systems have notable difficulties with these situations despite their pervasiveness in dialogue, as illustrated below.

User: I want to book a ticket to Newark, sorry New York

(TTT) Bot: There are multiple destinations, please select from the list below. [list includes Newark, NJ, Newark DE, and Newark NJ.]

The chatbot response does not handle the user's attempted self-correction. There are several possible answers that would be more appropriate here. First, the system could suggest New York, as the user intended. Second, the system could clarify if Newark or New York was the intended destination. This option is not exactly what the user requested, but would be a better response than only showing options for Newark. Previous methods that have been proposed for clarification generation (i.e., (Benotti and Blackburn, 2021; Kim et al., 2021b; Majumder et al., 2021)) could be employed to address this issue. Similarly, a pragmatically sensitive summary of this exchange would be *The user wanted to book a ticket to New York.*. It is unnecessary to capture the misstatement or its correction.

4.1.4 Semantic Imprecision

Anaphora is a common class of semantically imprecise words. To understand the intended meaning, people (or conversational AI systems) must determine the correct referent. This is conventionally achieved via linear order; given a pronoun, the correct referent is the most proximal noun (Carden, 1982). However, examination of discourse in use demonstrates this approach is flawed– pronouns regularly precede their referents or refer to remotely mentioned subjects (Carden, 1982). (Van Hoek, 1997) proposed that people use imaginary perspectives called *conceptual reference points* to view discourse and assign connections between items (Langacker, 1993). Use of these reference points allows people to assign pronouns to distal or upcoming referents.

In NLP, pronouns are often ignored (as in the case of removal with stop word lists (Nothman et al., 2019)) or addressed via replacement. While some work has found replacement to work well for discourse in a constrained environment (i.e., (Chen and Yang, 2021)), obvious errors can result when replacement is applied to naturalistic discourse, as illustrated in the example conversation summary below.

Interviewer: What do you do when you're annoyed?

Patient: [provides response]

Interviewer: Can you give me an example of that?

Patient: Uh, if someone ... annoys me I would let them know they're annoying me until they stop.

(TTT) Summary: What do you do when **they** are annoying?

The pragmatic failure of this example is subtle, but has clear implications for reader understanding. The interviewer asks the patient about their activities, specifically, what the patient does when *the patient* is annoyed. The patient provides a requested example, but uses the pronoun *they* to refer to a hypothetical annoying third party. The summarization model proposes a summary for the interviewer's question that replaces the second pronoun *you* with the pronoun *they*.

While the proposed summary is grammatically correct, it has two pragmatic failures. First, the summary does not include a referent for the pronoun. The meaning of *you* can be determined from the context– the interviewer is clearly speaking with a patient. The referent for *they* is unclear. The pragmatically appropriate summary in this case would be the original question.

Pronouns also convey point of view information (Van Hoek, 2010). For systems that require a complete description of the speaker's experience, incorrect pronoun assignment risks incorrect conclusions that lead to pragmatic failures. In the example above, the proposed summary changes the focus of the conversation. The interviewer's question was intended to gather information about the patient's behavior. The summary question shifts the focus of attention to the patient being annoyed.

Ambiguous pronouns violate the expected structure of summaries and therefore influence the coherence of reader's mental representations of text (Bransford and Johnson, 1972, 1973). Specifically, ambiguous pronouns reduce the coherence of reader's representations of discourse content (McNamara et al., 1996).

Speakers use point of view to account for their partners' perspectives (Lockridge and Brennan, 2002), prior expertise with a topic, (Isaacs and Clark, 1987), and perceptual abilities (Clark et al., 1983).

4.2 External Context

When speakers engage in conversation with others, they expect that their conversation partners are aware of salient features of their shared external context (Clark and Marshall, 1981). This includes both the physical environment and relevant background knowledge. Shared context controls detail in the exchange, allowing for the elimination of the obvious, and explicit emphasis on the non-obvious.

Previous research on user expectations for chatbots and conversational agents has found that users expect systems to account for external context and find it frustrating when they are unable to do so (Liao et al., 2016). Users want systems to be aware of their status and intentions and respond accordingly (Liao et al., 2016). Users create expectations of information that systems should know and want systems to use that information (Luger and Sellen, 2016). Including external context in the design of conversational AI applications, surely a challenging goal, will produce systems that are more consistent with users' expectations and more straightforward to use.

Empirical research has demonstrated that humans regularly utilize external context in conversation to provide the right amount of detail (Van der Henst and Sperber, 2004), account for a partner's expertise (Isaacs and Clark, 1987), and create references that partners will understand (Clark et al., 1983). Where previous work has incorporated context, the focus has largely been on conversational context. We suggest that this approach is insufficient– even if conversational context is represented perfectly, pragmatic failures will arise from a lack of appropriate awareness of relevant aspects of external context. In the following subsections, we discuss some of these pragmatic failures and relevant theoretical work.

4.2.1 Episodic Features

Speakers regularly use words and expressions like *today* that are semantically imprecise and are understood with reference to the current context (Levinson, 2011). These language functions are examined in the theoretical area of *deixis* (Levinson, 2011). They are tolerated in conversation when they are efficient to articulate and interpret. A previously un-grounded *there* for example, becomes tolerable when the speaker glances at the intended referent. For NLP applications, these expressions pose a semantic interpretation problem due to an impoverished representation of episodic conversation features.

Consistent with the notion of common ground, (Barwise and Perry, 1983) proposed that utterances require interpretation with respect to three situations. The *discourse situation* represents facts that someone might observe about the conversation. The *resource situation* includes the relationships between the speakers and facts known to all members. The *described situation* includes facts that could verify or falsify an utterance. This taxonomy reveals the scope of contextual features that are not always incorporated in the development of conversational AI applications.

Conversational AI applications often fail to incorporate not only external facts about the physical setting, like the user's location (represented in the *discourse situation*), but also semantic knowledge that all members of the conversation already know (represented in the *resource situation*). Examples of background knowledge include conceptual knowledge (Speer et al., 2017), domain specific knowledge (Gaur et al., 2018), attribute information (Zhang et al., 2016), commonsense knowledge (Davis and Marcus, 2015) and or information about the user.

This lack of appropriate awareness of external context generates several issues for conversational AI systems. First, relevance theory emphasizes that systems that lack awareness of relevant external context are unlikely to generate optimally relevant content for users. The voice assistant example below illustrates a pragmatic failure arising from a lack of this type of knowledge.

User: Is there a heat warning today?

(TTT) Assistant: I found this on the web *[Provides news article about heat wave in the UK when the user is in the US.]*

Pragmatically-appropriate assistant: Yes, there is a heat warning effect in [area] until [time].

As is typical of conventional conversation, the user does not specify their easily inferred location and the voice assistant fails to account for the user's location. The user does not receive the expected answer and must search elsewhere. A more appropriate answer is illustrated in the second response- it provides relevant information tailored to the user making the request. Moreover, a proper summary would actually *add* inferred content: the user wanted to know of heat warnings in [area]. Indeed, users create expectations about the kinds of information that conversational AI systems should have (i.e., location information from their profile) and would like systems to make use of that information (Luger and Sellen, 2016).

4.2.2 Conceptual Knowledge

Similarly, pragmatic failures can arise from an interaction between lack of episodic awareness and a lack of requisite semantic knowledge. The following voice assistant example illustrates this point.

User: Do I need gloves today?

(TTT) Assistant: Not much sun in the forecast today.

Pragmatically-appropriate assistant: The temperature is X degrees today.

In order to answer this question correctly, two types of knowledge are required. The first is knowledge of the current situation. The system must have an awareness of what day it is, where the user is located, and what the weather forecast is for that day in that location. Second, the system needs the conceptual knowledge about what gloves are and why the user might want to wear them (i.e., because it might be cold outside). This type of knowledge is often discussed under common sense reasoning (Davis and Marcus, 2015). The sample response successfully demonstrates awareness of the current situation. However, the system clearly lacks the requisite conceptual knowledge and gives an irrelevant answer about the amount of sun in the weather forecast. A more appropriate answer might include the low temperature for the day. A proper summary would also add content: The user asked if it was cold enough for gloves in [area] on [date]. For the user to determine if gloves are required, they must search elsewhere for the temperature or ask a more specific question. Indeed, users want systems to understand such intentions and respond accordingly (Liao et al., 2016).

4.2.3 Default Reasoning

Default reasoning addresses situations where available information is incomplete and conclusions need to be made based on what is generally true (Brewka, 2012). Conversation often contains instances that require default reasoning. Conversational AI applications need the ability to handle these instances. Consider the case of traffic delays illustrated in the example below.

User: I want to go to Cleveland, are there any traffic delays?

(TTT) Assistant: Getting directions to Cleveland [does not provide information about delays].

Pragmatically-appropriate assistant: Traveling to Cleveland will take 3 hours. There are no current delays. [Pulls up directions]

A pragmatically sensitive response to the query would acknowledge all sources of traffic delays likely to impact the user's trip. While the prototypical example of a traffic delay is construction, a pragmatically sensitive response would also account for other possible delays, such as a high probability snow storm. Similarly, a pragmatically sensitive summary of the exchange is: the user was concerned about travel delays from [area] to Cleveland and the system provided information about possible delays.

Default reasoning is a type of nonmonotonic reasoning. Unlike traditional logic, nonmonotonic reasoning addresses situations where new information can invalidate old conclusions (Brewka, 2012). In the above example, a pragmatically sensitive answer would account for the likelihood of a specific delay. It would not be pragmatically sensitive to provide a warning about a possible delay from a minor traffic slowdown several hours ahead.

4.2.4 Inconsistent Details

Another type of nonmonotonic reasoning that presents a challenge for conversational AI is reasoning given inconsistent details (Brewka, 2012). When two pieces of information are inconsistent, reasoners must determine what parts of the available information should be disregarded and what parts should be retained. Humans are generally able to resolve these sorts of inconsistencies (Johnson-laird et al., 2004). Conversational AI applications that lack these abilities will generate pragmatic errors, as illustrated in the voice assistant example below:

User: Remind me on Friday August 4th at 5:00 to order groceries. [Friday is August 5th, not August 4th]

(TTT) Assistant: Done [creates reminder for Thursday August 4th at 5:00]

Pragmatically-appropriate assistant: Did you mean Thursday August 4th or Friday August 5th?

To determine the correct action, the assistant needs to detect and then resolve the inconsistency of which piece of information the user intended, *Friday* or *the 4th*. A more appropriate response here would be to request clarification, as in the sample appropriate response. Failure to detect and resolve the inconsistent results in what is known as conversational breakdown (Ashktorab et al., 2019). A comparable summary would report the corrected date. Inconsistency is compounded where simultaneous activity occurs with conversation (such as in meetings). Comprehensive summaries in this area require reasoning about the state of the world based on dialogue.

To effectively resolve these inconsistencies, systems need the ability to detect inconsistent information and intervene. Previous work has developed methods for proposing clarification questions when conversational AI systems are unsure of the meaning of a user's utterance (i.e., (Benotti and Blackburn, 2021; Kim et al., 2021b; Majumder et al., 2021)). Similar methods could be employed to address situations where users provide inconsistent information as in the example. Moreover, these methods could address situations where other information indicates inconsistency, such as a user who asks a voice assistant to create a new calendar event that would overlap an existing event.

4.2.5 Domain Specificity

Lastly, external context pragmatic failures can arise from specific application environments where users have prior expertise with a given topic. These applications need communal common ground with the intended user (Clark, 1996) to support appropriate audience design (Bell, 1984). For example, it is appropriate to define new anatomy terminology in lecture summaries or automated tutoring systems. This same terminology should not be defined in a summary of a meeting between doctors discussing the statuses of current patients. Similarly, virtual assistants, ASR systems, content filters, and other applications need an awareness of domain content when developed for domain-specific applications. One humorous real world example is a profanity filter for a virtual archaeology conference that banned the word *bone* (Ferreira, 2020).

5 Discussion

We have demonstrated that the challenges to automated conversation summary, what we have termed discourse over discourse, are symptomatic of a more fundamental, general problem regarding the absence of attention to the role of pragmatics in automated conversation. We proposed that the pragmatic failures of conversational AI systems are captured by *relevance theory* (Wilson and Sperber, 2013). Relevance suggests two key issues for conversational AI systems: preservation of meaning and awareness of external context.

While previous NLP work has examined pragmatic issues separately, in language models (Pandia et al., 2021; Ettinger, 2020; Gubelmann and Handschuh, 2022; Wang et al., 2021), downstream tasks (Nie et al., 2020; Schüz and Zarrieß, 2021; Zhang et al., 2022a), dialogue systems and conversational models (Bao et al., 2022; Kim et al., 2020, 2021a; Nath, 2020; Wu and Ong, 2021), our integrative approach is intended to distill and taxonomize recurrent foundational themes to motivate a theoretical framework and coordinated research efforts. Similarly, we suggest that a theoretical framework will facilitate response to the large body of applied work on human expectations for conversational AI applications (Ashktorab et al., 2019; Liao et al., 2016; Luger and Sellen, 2016; Zamora, 2017). We aim to provide such a framework by integrating these issues with theoretical and empirical work in pragmatics.

5.1 Limitations and Ethical Considerations

This class of work has several important limitations and ethical concerns. First, this work inherits privacy concerns common to these types of applications. Many of the features represented in external context are not necessarily directly accessible via the semantic content in a conversation (i.e., user location). While some users want systems to use this information (Luger and Sellen, 2016), others may not. Systems should clearly illustrate needed information, intended use and storage. Users should be able to easily and accessibly customize what information they share. Moreover, it is important to avoid creating systems that provide a sub-optimal user experience for users who do not want to share information. (Zuboff, 2020) points out that some products are essentially not usable without agreeing to the product's data use policy. Creating systems that can pose clarification questions is one way of addressing this issue. If a user asks a question that requires location information they have not shared, the system could ask if there is a specific location the user would like a response for.

Second, our position could be interpreted as endorsing the development of deep learning models with high monetary and energy costs (Strubell et al., 2019). We point out that many of the issues we raise could be addressed with approaches that utilize pre-existing external knowledge (Valiant, 2006), such as knowledge graphs (Miller, 1995; Speer et al., 2017) or lexicons (Gaur et al., 2018; Sheth et al., 2005), that can reduce the need to acquire this information through deep learning.

6 Conclusion

Several pragmatic challenges recur across current conversational AI applications. Drawing on relevant theoretical work in linguistics and psycholinguistics, we examine each of these challenges in detail. We illustrate our points with examples that are syntactically correct, but have clear pragmatic deficiencies and integrate our observations with HCI research that has examined user expectations and frustrations surrounding current conversational AI applications. These results contribute to a better understanding of current pragmatic challenges and suggest areas for improvement. Two of these needs most salient in our review are better connection to general knowledge and the external environment. Contrary to the notion of summaries as simplified versions of discourse, we observe that comprehensive summaries often require adding content to clearly to relevant general knowledge and aspects of the external environment. Future work can examine possible approaches to developing systems that can address these challenges and better meet the pragmatic expectations of users.

References

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow Scotland Uk. ACM.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842, Dublin, Ireland. Association for Computational Linguistics.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. MIT Press, Cambridge, MA.

Isabel L. Beck, Margaret G. McKeown, Gale M. Sinatra, and Jane A. Loxterman. 1991. Revising Social Studies Text from a Text-Processing Perspective: Evidence of Improved Comprehensibility. *Reading Research Quarterly*, 26(3):251–276. Publisher: [Wiley, International Reading Association].

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

Luciana Benotti and Patrick Blackburn. 2021. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why People Use Chatbots. In *Internet Science*, Lecture Notes in Computer Science, pages 377–392, Cham. Springer International Publishing.

John D. Bransford and Marcia K. Johnson. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6):717–726.

John D Bransford and Marcia K Johnson. 1973. Considerations of some problems of comprehension. In *Visual Information Processing*, pages 383–438. Academic Press.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning Memory and Cognition*, 22(6):1482–1493.

Gerhard Brewka. 2012. Introduction. In *Nonmonotonic Reasoning: Logical Foundations of Commonsense*. Cambridge University Press, Cambridge, MA.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Herman Buschke and Aron H. Schaier. 1979. Memory units, ideas, and propositions in semantic remembering. *Journal of Verbal Learning and Verbal Behavior*, 18(5):549–563.

Guy Carden. 1982. Backwards anaphora in discourse context. *Journal of Linguistics*, 18(2):361–387.

Rollo Carpenter/Existor. 2022. Cleverbot.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Herbert H. Clark. 1979. Responding to indirect speech acts. *Cognitive Psychology*, 11(4):430–477.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

Herbert H Clark and Catherine R Marshall. 1981. Definite reference and mutual knowledge. In A.K. Joshi, B.L. Webber, and I.A. Sag, editors, *Elements of discourse understanding*, pages 10–63. Cambridge University Press, Cambridge. ISSN: 0749596X.

Herbert H Clark and Edward F Schaefer. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2(1):19–41.

Herbert H. Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2):245–258.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

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. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Becky Ferreira. 2020. A Profanity Filter Banned the Word 'Bone' at a Paleontology Conference.

Kenneth I. Forster. 1970. Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics*, 8(4):215–221.

Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Manas Gaur, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L. Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, and Amit Sheth. 2021. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8(5):1–19.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. " Let Me Tell You About Your Mental Health!" Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 753–762.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4769–4772. ISSN: 2379-190X.

Ernest T. Goetz, Richard C. Anderson, and Diane L. Schallert. 1981. The representation of sentences in memory. *Journal of Verbal Learning and Verbal Behavior*, 20(4):369–385.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742.

Arthur C. Graesser, Nicholas L. Hoffman, and Leslie F. Clark. 1980. Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior*, 19(2):135–151.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of LREC 2014 May*, pages 3123–3128.

H.P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics 3: Speech acts*, pages 41–58. Academic Press, New York.

Reto Gubelmann and Siegfried Handschuh. 2022. Context matters: A pragmatic study of PLMs' negation understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4602–4621, Dublin, Ireland. Association for Computational Linguistics.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep Neural

Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Ellen A. Isaacs and Herbert H. Clark. 1987. References in Conversation Between Experts and Novices. *Journal of Experimental Psychology: General*, 116(1):26–37.

P. N. Johnson-laird, Vittorio Girotto, and Paolo Legrenzi. 2004. Reasoning from inconsistency to consistency. *Psychological Review*, 111(3):640–661.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021a. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joo-Kyung Kim, Guoyin Wang, Sungjin Lee, and Young-Bum Kim. 2021b. Deciding whether to ask clarifying questions in large-scale spoken language understanding. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 869–876, Cartagena, Colombia. IEEE.

Walter Kintsch. 1974. *The representation of meaning in memory*. The representation of meaning in memory. Lawrence Erlbaum, Oxford, England. Pages: vii, 279.

Walter Kintsch and Janice Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5(3):257–274.

Ronald W Langacker. 1993. Reference-point constructions. *Cognitive Linguistics*, 4(1):1–38.

Stephen C. Levinson. 2011. Deixis. In Laurence R. Horn and Gregory Ward, editors, *The handbook of pragmatics*, pages 97–121. Blackwell Publishing Ltd.

Piji Li, Wai Lam, Lidong Bing, Weiwei Guo, and Hang Li. 2017. Cascaded attention based unsupervised information distillation for compressive summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2081–2090, Copenhagen, Denmark. Association for Computational Linguistics.

Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 264–275, Brisbane QLD Australia. ACM.

Calion B Lockridge and Susan E Brennan. 2002. Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin and Review*, 9(3):550–557.

Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5286–5297, San Jose California USA. ACM.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Danielle S McNamara, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1):1–43. Publisher: Taylor & Francis.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Anindita Nath. 2020. Towards naturally responsive spoken dialog systems by modelling pragmatic-prosody correlations of discourse markers. *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 128–129. ISBN: 9781450375139.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.

Joel Nothman, Hanmin Qin, and Roman Yurchak. 2019. Stop Word Lists in Free Open-source Software Packages. In *Proceedings of Workshop for NLP Open Source Software*, pages 7–12.

L Nyberg, A R McIntosh, R Cabeza, R Habib, S Houle, and E Tulving. 1996. General and specific brain regions involved in encoding and retrieval of events: what, where, and when. *Proceedings of the National Academy of Sciences*, 93(20):11280–11285.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada. ACM.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Simeon Schüz and Sina Zarrieß. 2021. Decoupling pragmatics: Discriminative decoding for referring expression generation. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 47–52, Gothenburg, Sweden. Association for Computational Linguistics.

Amit Sheth, Boanerges Aleman-Meza, I Budak Arpinar, Clemens Bertram, Yashodhan Warke, Cartic Ramakrishanan, Chris Halaschek, Kemafar Anyanwu, David Avant, F Sena Arpinar, and others. 2005. Semantic association identification and knowledge discovery for national security applications. *Journal of Database Management (JDM)*, 16(1):33–53. Publisher: IGI Global.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *31st AAAI conference on artificial intelligence*, pages 4444–4451.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press, Cambridge, Massachusetts.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27.

Endel Tulving. 1972. Episodic and semantic memory. In *Organization of memory*, pages xiii, 423–xiii, 423. Academic Press, Oxford, England.

Leslie G Valiant. 2006. Knowledge infusion. *AAAI*, 6.

Jean-Baptiste Van der Henst and Dan Sperber. 2004. Testing the Cognitive and Communicative Principles of Relevance. In Ira A. Noveck and Dan Sperber, editors, *Experimental Pragmatics*, Palgrave Studies in Pragmatics, Language and Cognition, pages 141–171. Palgrave Macmillan UK, London.

Teun A. van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press, New York. ISSN: 00978507.

Karen Van Hoek. 1997. Backwards anaphora as a constructional category. *Functions of Language*, 4(1):47–82.

Karen Van Hoek. 2010. Pronominal Anaphora. In Dirk Geeraerts and Hubert Cuyckens, editors, *The Oxford Handbook of Cognitive Linguistics*, pages 1–25. Oxford University Press.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219, Online. Association for Computational Linguistics.

Rose Wang, Julia White, Jesse Mu, and Noah Goodman. 2021. Calibrate your listeners! robust communication-based training for pragmatic speakers. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 977–984, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deirdre Wilson and Dan Sperber. 2013. Relevance theory. *Routledge Companion to Philosophy of Language*, pages 163–176. ISBN: 9781136594083.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Zhengxuan Wu and Desmond C. Ong. 2021. Pragmatically informative color generation by grounding contextual modifiers. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 438–445, Online. Association for Computational Linguistics.

Jennifer Zamora. 2017. I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pages 253–260, New York, NY, USA. Association for Computing Machinery.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022a. Improving the adversarial robustness of NLP models by information bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3588–3598, Dublin, Ireland. Association for Computational Linguistics.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *Proceedings of the 22nd {ACM SIGKDD} International Conference on Knowledge Discovery and Data Mining*, pages 353–362, San Francisco California USA. ACM.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [cs].

Shoshana Zuboff. 2020. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, 1st edition. PublicAffairs.