

# Chapter 9

## Intelligent Agents and Dialog Systems



Timothy Bickmore and Byron Wallace

**After reading this chapter, you should know the answers to these questions:**

- What is a dialog system and how can it be used in patient- and consumer-facing systems in medicine?
- What are the main approaches to the implementation of dialog systems? What are the limitations of these approaches?
- How are dialog systems evaluated?
- What are some of the safety issues in fielding patient- and consumer-facing dialog systems in medicine?

### Introduction to Dialog Systems

People most commonly communicate with each other not in isolated utterances, but in interleaved sequences of utterances wrapped in ritualized behavior that we colloquially refer to as conversations. Developing natural language interfaces that can move beyond single transactions of user query/system response to fully engage users in conversation would benefit a variety of applications. At a minimum, once the information that needs to be exchanged extends beyond that which can be expressed in a single utterance, dialog becomes imperative. Beyond this, dialog is essential for performing tasks that require multiple natural language exchanges with a user in a coherent manner, as for example in a series of questions and responses to automate an interactive, incremental differential diagnosis. Certainly, the emulation

---

T. Bickmore (✉) · B. Wallace  
Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA  
e-mail: [t.bickmore@northeastern.edu](mailto:t.bickmore@northeastern.edu); [b.wallace@northeastern.edu](mailto:b.wallace@northeastern.edu)

© The Author(s), under exclusive license to Springer Nature  
Switzerland AG 2022

T. A. Cohen et al. (eds.), *Intelligent Systems in Medicine and Health*, Cognitive Informatics in Biomedicine and Healthcare,  
[https://doi.org/10.1007/978-3-031-09108-7\\_9](https://doi.org/10.1007/978-3-031-09108-7_9)

of any kind of counseling session or interview to produce automated patient-facing health education systems requires complex goal-oriented dialog management that spans many interleaved patient and system messages. In addition, interleaved sequences of messages allow a listener to confirm understanding or request clarification of information provided (a process referred to as “grounding”). Only in dialog can a conversational task (e.g., diagnosis or health counseling) be dynamically decomposed into sub-tasks in a coherent manner.

To ground our discussion, Fig. 9.1 shows an excerpt of a dialog between a study nurse and a patient about informed consent for an oncology clinical trial. There are several interesting things to note for those who have not studied natural conversations before. First, spontaneous conversation is full of *disfluencies*: there are very few grammatically complete and correct sentences in spontaneous conversation, and the use of “filler words” such as “um” (as in line #11) is very common. Second, conversational turns can span a single word to many sentences in duration. Third, a great deal of conversation is spent establishing mutual understanding of what was said: the patient feedback at lines #2 and #10, and the patient query at line #14 all serve exclusively to ensure that both parties understand each other, at least well enough for the purpose at hand. Only one person can talk at a time in conversation, and people are generally very good at coordinating their use of the speech channel, but overlaps, pauses (as in line #11) and interruptions (such as in line #15) are common. Finally, conversation typically makes extensive use of “deixis”, which is a reference to the immediate physical context or to what was said before (for example, line #1 refers to the current day, line #9 refers to the consent form that is being handed to the patient). Designing automated dialog systems that can participate in these kinds of conversations, for example taking the role of the study nurse here to

1. Nurse: So, today I am going to talk to you about a research study that your doctor has stated that might be a good option for you.
2. Patient: OK.
3. Nurse: Have you had any experience with research before?
4. Patient: I have done a couple of research before. Nothing like this, just questionnaires on how I was treated as a patient. You know, small little things like 25 dollars.
5. Nurse: All right, but no treatment. You never took any drugs or anything like that?
6. Patient: No.
7. Nurse: Have you ever been treated for cancers in the past?
8. Patient: This is the first time.
9. Nurse: First time. OK, what I am going to do is give you this packet. And this is something that we give to all of our patients. This is the consent form.
10. Patient: OK.
11. Nurse: Ummmm. So.....(Pause)... This drug that Doctor Smith said that you may be eligible for is a drug that is kind of daughter or son of Thalidomide. Have you ever heard of Thalidomide?
12. Patient: No, never heard of it.
13. Nurse: Thalidomide was used back in the 50's and 60's for nausea in European women and...
14. Patient: (interrupting) Feeling sick?
15. Nurse: Yeah, feeling sick. And, they took this drug that does wonders for the nausea but unfortunately because they were pregnant during that it caused side-effects. ...

**Fig. 9.1** Excerpt of nurse-patient dialog for administration of oncology clinical trial informed consent

automate administration of informed consent, represents an aspirational goal for dialog systems researchers. However, the state of the art is quite far from achieving this level of performance.

Chapter 7 introduced natural language processing (NLP). In this chapter we review the state of the art in dialog systems, a sub-field of NLP, including text-based chatbots, speech-based conversational assistants, and multimodal embodied conversational agents that simulate face-to-face conversation, for both provider- and patient-facing biomedical applications.

## *Definitions and Scope*

**Dialog** has been defined as a conversational exchange between two or more entities. For the purposes of this chapter, we will be concerned with communicative exchanges between a human (health professional, patient, or consumer) and an automated system in which messages are in textual or spoken natural language. This system may also be augmented with additional information such as the non-verbal behavior used by humans in face-to-face conversation (hand gestures, facial displays, eye gaze, etc.). We refer to an isolated message from one entity within a dialog as an **utterance**. While a dialog can consist of a single utterance, we are primarily concerned with dialogue in which several utterances from two entities are interleaved in order to accomplish some task. **Discourse** is a generalization of dialog that also includes the study of written text comprising multiple sentences.

Discourse theory is generally concerned with how multiple utterances fit together to specify meaning. Theories of discourse generally assume that discourses are composed of **discourse segments** (consisting of one or more adjacent utterances), organized according to a set of rules. Beyond this, however, discourse theories vary widely in how they define discourse segments and the nature of the inter-segment relationships. Some define these relationships to be a function of surface structure (e.g., based on categories of utterance function, such as *request* or *inform*, called “speech acts” [1]), while others posit that these relationships must be a function of the intentions (plans and goals) of the individuals engaged in conversation [2, 3]. In addition, researchers developing computational models of discourse and dialog have included a number of other constructs in their representation of discourse context, including: entities previously mentioned in the conversation; topics currently being discussed (e.g., “questions under discussion” [4]); and information structure, which indicates which parts of utterances contribute new information to the conversation as opposed to those parts that serve mainly to tie new contributions back to earlier conversation [5].

Discourse theory also seeks to provide accounts of a wide range of phenomena that occur in naturally-occurring dialog, including: mechanisms for conversation initiation, termination, maintenance and turn-taking; interruptions; speech

intonation (used to convey a range of information about discourse context [6]); discourse markers (words or phrases like “anyway” that signal changes in discourse context [7]); discourse ellipsis (omission of a syntactically required phrase when the content can be inferred from discourse context); grounding (how speaker and listener negotiate and confirm the meaning of utterances through signals such as head-nods and paraverbals such as “uh huh” [8]); and indirect speech acts (e.g., when a speaker says “do you have the time?” they want to know the time rather than simply wanting to know whether the hearer knows the time or not [9]).

### Box 9.1 Definition

**Dialog systems** are computational artifacts designed to engage humans in dialog, as defined above. **Intelligent agents** are autonomous, goal-directed computational artifacts. **Conversational agents** are intelligent agents that converse with humans via a dialog system interface. **Conversational assistants** are conversational agents that use speech input and output to perform a wide range of tasks, as exemplified by the now ubiquitous Siri, Amazon Alexa, and Google Home products.

*Embodied Conversational Agents* (ECAs) are conversational agents that include the ability to use human-like conversational nonverbal behavior in their dialog (Fig. 9.2). ECAs are animated humanoid computer-based characters that use speech, eye gaze, hand gesture, facial expression and other nonverbal modalities to emulate the experience of human face-to-face conversation with their users [10]. Such agents can provide a “virtual consultation” with a simulated health provider, offering a natural and accessible source of information for patients. These agents represent one form of multimodal dialog system, in which the nonverbal modalities are recognized and produced in addition to accompanying text or speech, to more fully understand the user’s communicative intent and to better express system

**Fig. 9.2** Embodied conversational agent for patient education at hospital discharge



output. In addition to carrying additional factual information, nonverbal behavior is also used in face-to-face conversation to regulate the interaction structure itself, for example, gaze and intonation to regulate turn-taking behavior, and body position and orientation to regulate conversation initiation and termination. Nonverbal behavior is also particularly effective for conveying affective and relational cues that may be important for establishing patient trust in and working alliance with the ECA [11].

### ***What's Hard About Getting Machines to Engage in Spontaneous Human Conversation?***

People unconsciously leverage a complex set of processes to make conversation work, most of which are entirely automatic and unconscious. They assume that an entity that engages them in what appears to be natural language dialog has these abilities until they discover their limitations. Several of these processes, such as conversation initiation and termination, and turn-taking and grounding, were mentioned above. Additional examples include: deixis, referring to something in the speakers' mutual context (object, time, location, social relationship) in language; anaphoric or cataphoric references (referring to something said earlier or later in the dialog); and conversational framing [12] or layering [13] in which different styles or genres of talk are used to change how utterances are interpreted (e.g., symptom inquiry by a third party made within social chat storytelling occurring within the context of a clinical interview). There are many more conversational processes and linguistic phenomena that together make the seemingly effortless task of a water cooler conversation seem miraculous upon close inspection.

Fortunately, most of these conversational processes can be “compiled out” by tightly constraining what a user is allowed to do, or by greatly lowering their expectations. System-initiated dialog that rigidly walks a user through a series of steps generally avoids the need to engage in many of these processes. Similarly, a system that engages a patient in scripted greeting and small talk at the start of a health education session does not need a computational model of conversational frames. Agents that only provide responses to single utterance user queries (such as popular conversational assistants like Siri) have trained users to not expect any conversational behavior beyond these simple exchanges.

### ***Machine Learning and Dialog Systems***

In the research community, the dominant modern approach to dialog systems is now based on machine learning (ML; see Chap. 6). Learning-based approaches to dialog permit flexibility and avoid the need for exhaustive manual engineering of rules.

ML-based approaches to dialog have yielded strong empirical performance, although measuring this is a challenge ([14]; see the section on “Automated Metrics for End-to-End Architectures”). However, building such systems requires *training data* (i.e., example conversations) from which to learn, which may not be available in all domains, and can be prohibitively expensive to collect. Moreover, it is difficult to *control* the outputs of machine learning models, and so deploying such systems in the context of healthcare applications may be a risky endeavor.

## History of Dialog Systems in Healthcare

Chapter 2 reviewed the history of AI in medicine; here we focus on the development of dialog systems specifically. One of the very earliest dialog systems developed was produced as a demonstration of a patient-facing psychotherapy counseling agent (see Chap. 2). The ELIZA system was developed to simulate the behavior of a Rogerian psychotherapist, in which the patient and the computer exchanged typed text messages [15]. Although ELIZA was not intended to be used for actual therapy, similar systems have been proven effective for therapy in which the system is essentially prompting a patient to think aloud and work through his or her own problems [16]. An example conversation with ELIZA is shown in Fig. 9.3.

Colby developed an ELIZA-like system that was designed to use Cognitive Behavioral Therapy to treat individuals with depression. In addition to providing typed text counseling with patients, the system provided text-based educational materials about depression [17]. These systems are characterized by system responses that are only coherent with the immediately preceding user utterance, implemented using pattern-response rules that are matched to the user input with regular expressions, and template-based text generation of system responses. They also use a variety of techniques to maintain the illusion of coherent dialog, including: maintaining system-initiated dialog, having most system outputs prompt the user with open-ended questions; relying on the user’s sense-making ability to infer

1. ELIZA: Hello, I am Eliza.
2. User: Hi.
3. ELIZA: How are you today.. What would you like to discuss?
4. User: I've been feeling a little depressed lately.
5. ELIZA: Can you elaborate on that?
6. User: I'm just not myself. I find it hard to get motivated.
7. ELIZA: Did you come to me because you are just not yourself. You find it hard to get motivated?
8. User: Yes, that's right.
9. ELIZA: I see.

**Fig. 9.3** Example Conversation with ELIZA

coherent explanations for the system’s outputs; and reflecting the user’s inputs back to them with minor wording changes in order to give the illusion of understanding what the user is saying. This approach to dialog system implementation is widely used in “chatbots” deployed on the web for entertainment, marketing and sales applications, and has given rise to an open standard chatbot implementation language (AIML [18]).

Development dating from 1964 was also conducted on systems that could collect a medical history from patients [19]. Unlike ELIZA, these systems conducted system-initiated dialog only, asking patients a series of questions with highly-constrained patient input (mostly YES/NO questions) to drive branching logic. Research and development of these systems has continued, and some commercial tools are available, although they still have not attained wide use in clinical practice [20].

Some of the earliest work in physician-facing medical expert systems used system-initiated dialogue to interact with providers for decision support. MYCIN was an early rule-based expert system that identified bacteria causing an infection and recommended antibiotics [21] (see Chap. 2). It was designed to interact with physicians by asking a series of very constrained questions requiring one- or two-word responses. In fact, it was a desire to avoid having to implement natural language understanding that led to the use of MYCIN’s core backward-chaining diagnostic algorithm.

By using a backward-chained approach, MYCIN controlled the dialogue and therefore could ask specific questions that generally required one- or two-word answers. ([21], p. 601)

MYCIN (and derivative projects) used various text generation techniques to produce their final output case summaries.

The sections on “Example Patient- and Consumer-facing Dialog Systems” and “Example Provider-facing Dialog Systems” provide more recent examples of patient- and provider-facing medical dialog systems.

In the last decade, deep neural network-based methods trained on massive corpora have come to dominate Natural Language Processing (NLP) [22]. These methods have enabled highly accurate automatic speech transcription tasks [23] and improved NLP system performance across a variety of problems, including building conversational agents [24].

One means for building dialog systems entails specifying models that map user input utterances directly to output utterances (“end-to-end” systems). This can yield strong performance with respect to the fluency of outputs, but such systems can struggle to maintain coherence throughout a dialog [25]. While such text-to-text models have been used in the context of task-oriented dialog systems [26, 27], they may be more suitable to “general domain” conversational agents—i.e., general “chatbots”—as such models are not naturally amenable to guiding “goal-based”

dialog. This reflects the myopic optimization strategy used to estimate model parameters: Typically one aims to find parameters that make the model as likely as possible to produce the words comprising response utterances in the observed training data. This optimization criterion does not explicitly encode higher-order conversational goals, which likely require explicit planning to achieve.

A common strategy to address this problem is to decompose dialog systems into independent modules, to be trained separately and combined in a pipeline. For example, one module might process user utterances, a second might then decide on an action to take, and a third might then generate a response, conditioned on this. Developing “end-to-end” methods that permit joint optimization of all components necessary for goal-based dialog is an active area of research [28]. We discuss archetypal modern machine learning models in the section on “Neural Network Methods and End-to-End Architectures”.

## Dialog System Technology

### *Classic Symbolic Pipeline Architectures*

Historically, dialog systems have been developed using a pipeline architecture, in which a user utterance is incrementally transformed into a representation that the core agent logic can provide a response to, followed by another series of processing stages to render the system output. These stages can include Automated Speech Recognition (ASR), multimodal integration, utterance understanding, dialog management, natural language generation, multimodal generation, and Text-To-Speech (TTS). Approaches to dialog management include **finite-state automata**, **frames**, and plan-based frameworks (Fig. 9.4).

**Automated Speech Recognition** (ASR) is responsible for transcribing the users’ speech input into one or more text representations. Speech recognition has improved significantly from single-speaker digit recognition systems in 1952 [29] to speaker-independent continuous speech recognition systems based on deep neural networks [30]. Currently, several open source ASR engines such as Pocketsphinx [31], Kaldi [32], and HTK [33] are available, but accurate speech recognition can require substantial processing power which cloud based services such as IBM Watson [34], and the Google cloud platform [35] provide. Although recent systems have achieved around 5% word error rates [36, 37], there are still some doubts regarding the use of ASR in applications such as medical documentation [38]. Goss et al [39] reported that 71% of notes dictated by emergency physicians using ASR contained errors, and 15% contained critical errors.

A **Natural Language Understanding** (NLU) module extracts a semantic representation of the user’s utterance, which can then be used by the dialog manager to generate a system response. State-of-the-art statistical NLU systems often contain three main components: domain detection, intent detection, and slot tagging [40].



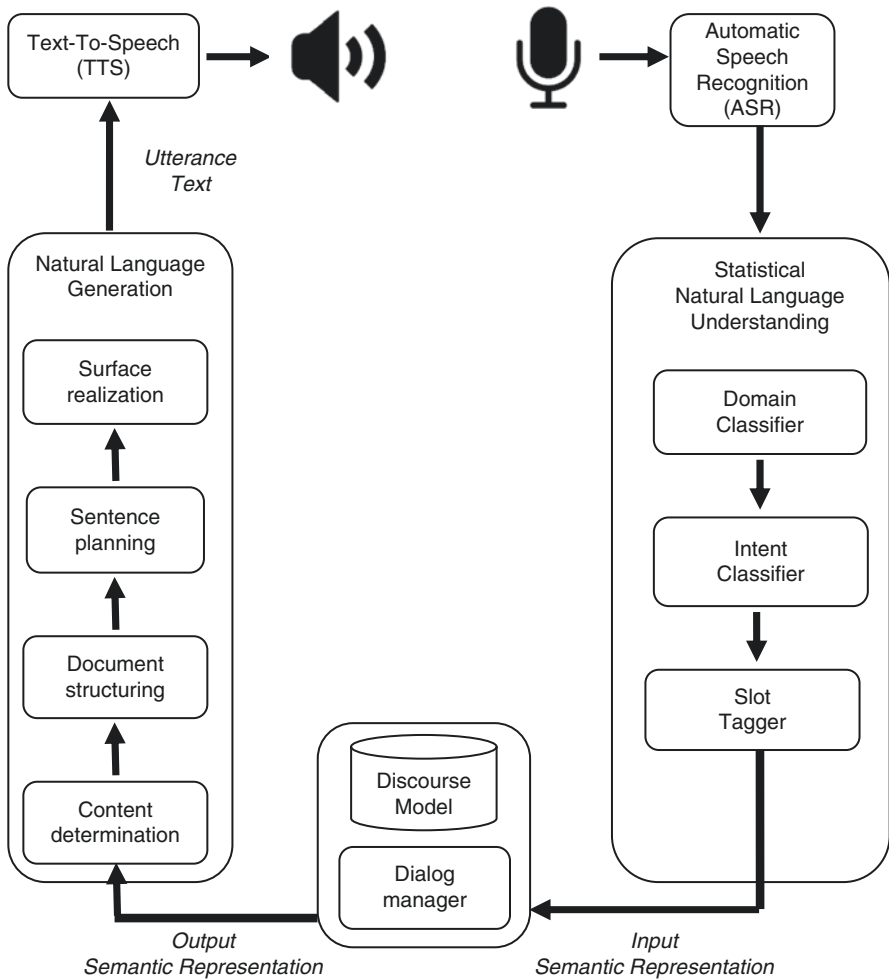


Fig. 9.4 Pipeline dialog system architecture

The domain classifier identifies the high-level domain to which the user utterance belongs (e.g., symptoms, medications, or educational content). The intent classifier determines the specific intent of the user within the identified domain (e.g. *report\_finding*). Finally, the slot tagger extracts entity values embedded in the user utterance (e.g. *syndrome\_name* or *severity\_level*). NLU is one of the most complex tasks in dialog systems for several reasons. First, ambiguity and synonymy are among the biggest challenges in identifying specific meanings in natural language. Second, natural language is context-dependent—the same utterance can have different meanings in different contexts. Third, spontaneous speech is often noisy with disfluencies (e.g., filled pauses, repairs, restarts).

Dialog management is most typically implemented using finite state machines or layers of finite state machines, also referred to as hierarchical transition networks, particularly for applications in which the system maintains the conversational initiative. In these systems, states typically represent system utterances and branches to next states are made in response to user responses. Layers in the hierarchy can be used to represent discourse segments, for example to satisfy a particular conversational goal. Dialog managers can also be frame-based, in which a current “frame” is used to guide the conversation by asking users for information to fill slots until enough information has been gathered for the system to take an action.

More advanced approaches to dialog management involve the explicit representation of user and system plans and goals, which is required to manage conversational phenomena such as: mixed-initiative dialog, in which either the user or the system can take control of the conversation at any time; proper handling of interruptions and requests for clarifications; and indirect speech acts. Flexibly handling these phenomena requires representing and reasoning about the intentions that underlie system and user utterances, inferring the user’s goals and task plan, and dynamically synthesizing the system’s task plan. Inferring a user’s goals and task plan is necessary because, as exemplified by indirect speech acts, people’s utterances do not always correspond directly to their communicative intent (e.g., as in “Do you have the time?”). Thus, plan-based theories of communicative action and dialog assume that the speaker’s speech acts are part of a plan, and the listener’s task is to infer it and respond appropriately to the underlying plan, rather than just to the utterance [41]. Synthesizing system task plans, including communicative and other actions, is necessary in complex applications in which all possible conversational contingencies (and their possible orderings) cannot be anticipated and scripted, but must be addressed in an incremental, reactive manner.

Dynamic planning and plan inference can be computationally very complex and difficult to develop, and thus have not been used much to date in fielded health dialog systems. However, they remain active areas of research, and a handful of health dialog systems that use these techniques have been developed for the application of clinical guidelines [42], for the automatic generation of reminders for older adults with cognitive impairment [43], for medication advice [44], and for diet promotion [45].

One research project used a task decomposition planning formalism to drive health behavior change counseling dialog for exercise and diet promotion [46]. This formalism was based on the Shared Plans theory [47, 48], in which dialog is viewed as a collaboration in which participants coordinate their action towards achieving a shared goal. Discourse segments are defined by the sequence of sub-goals or atomic actions in a recipe that serve to elaborate a particular goal, and the only meaningful relationships among discourse segments are elaboration (goal expansion) and ordering of goals and actions. Figure 9.5 shows a portion of the plan tree for an exercise promotion dialog. Plan fragments that elaborate dialog goals into subgoals and atomic actions are referred to as recipes and are represented in ANSI/CEA-2018 [49] (ANSI/CEA-2018 provides a standard declarative representation for tasks that can be decomposed in this manner). Figure 9.6 shows a portion of a high-level

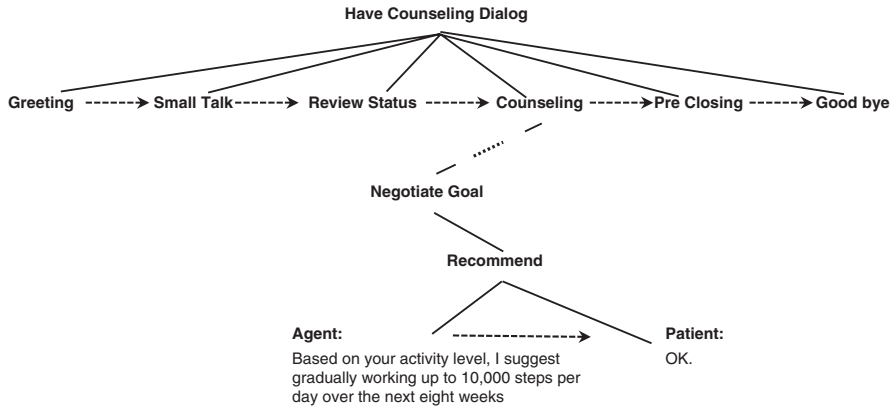


Fig. 9.5 Plan tree fragment for exercise promotion counseling dialog

Fig. 9.6 Example pseudocode for a high-level recipe and a low-level dialogue specification

<p><b>task</b> Negotiate</p> <p><b>input parameters:</b> behavior, target</p> <p><b>outputs:</b> goal</p> <p><b>steps:</b> 1. ComputeDesired 2. Recommend 3. Followup 4. Confirm</p>
<p><b>task</b> Recommend</p> <p><b>input parameters:</b> behavior, target, desired</p> <p><b>outputs:</b> response</p> <p><b>precondition:</b> (behavior='exercise' and target='long_term_goal')</p> <p><b>adjacency pair:</b></p> <p><b>agent utterance:</b></p> <p>Based on your activity level, I suggest gradually working up to &lt;desired&gt; steps per day over the next eight weeks</p> <p>...</p>

recipe for behavior goal negotiation, and an example of an atomic dialogue turn that elaborates the “Recommend” subgoal for negotiating long-term exercise goals. The run-time planning system (based on the COLLAGEN collaborative dialog system [50]) starts with a top-level goal to have a counseling dialog, then incrementally elaborates the goal using recipes until atomic utterances are produced. This process results in a plan tree in which the root is the initial goal and the leaves are the utterances produced by the agent and the user (Fig. 9.5). The planning process proceeds without backtracking, i.e., elaborations are never undone once they are added to the dialog plan tree.

The most common approach to symbolic **Natural Language Generation** (NLG) is template-based text generation, in which an output utterance is represented as a string annotated with variables whose values are determined at runtime [51]. While relatively simple and straightforward, this approach does not offer much flexibility or expressivity. In the most general case, text generation can involve word-by-word synthesis of utterances based on a grammar and dictionary, discourse context and world knowledge, and is itself decomposed into another pipeline of processing stages [51] (Fig. 9.4).

Content determination is the first stage, and involves deciding what information should be communicated in the output, beyond that dictated by the dialog manager. *Document structuring* decides how chunks of text should be grouped together in one or more output utterances and how they should be related in rhetorical terms. *Sentence planning* involves: selection of the specific words or other linguistic resources that should be used to express the selected content; deciding what expressions should be used to refer to entities; and deciding how structures created by document planning should be mapped onto linguistic structures, such as utterances or conversational turns.

The last step of NLG, referred to as **surface realization**, involves turning the internal representations produced during sentence planning into the text of one or more utterances. Research has also been conducted into generation of multi-modal system outputs (speech or text plus accompanying nonverbal behavior for an ECA, or graphics to help illustrate a concept to be conveyed) although, as with multi-modal input understanding, this has not been used widely in health dialog systems to date.

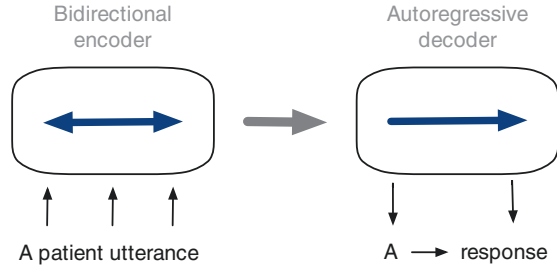
Finally, **Text-To-Speech** (TTS) involves the conversion of utterance text into an acoustic signal. TTS is now a very mature technology and the quality and naturalness has improved significantly over the last decade, producing understandable speech for a wide range of languages. Speech Synthesis Markup Language (SSML) enables the annotation of utterance text with tags that can manipulate speed, pitch, volume, and other aspects of prosody to produce more expressive speech [52].

### *Neural Network Methods and End-to-End Architectures*

In the past decade, neural networks have emerged as the dominant model class for natural language processing (NLP) [22], as they have become the dominant machine learning formalism for many areas of modeling in medicine (see Chap. 1). Neural network-based NLP has in turn given rise to neural conversational models [25, 53–55]. Departing from the classical symbolic approaches reviewed above, neural models represent utterances as dense, continuous vectors (i.e., learned representations). Neural language models [56] are typically used in such architectures to generate responses conditioned on a representation of context, e.g., the most recent utterance.

Completely “end-to-end” systems forego explicit planning and learn to map directly from an input to an output utterance via a deep neural network [57].

**Fig. 9.7** High-level schematic of a Seq2Seq model for dialog generation



*Sequence-to-sequence* architectures are currently dominant neural models for dialog; these use neural network modules to map input to output texts. This is typically accomplished with an *encoder-decoder* architecture. The encoder learns to compress inputs into a dense representation; this is then passed onto the decoder, which is responsible for conditionally generating a response. Because the decoder must generate text, it is typically defined as an auto-regressive (conditional) language model, which is to say that it generates outputs one word at a time in a left-to-right fashion. Figure 9.7 provides a high-level schematic of this approach.

Simple sequence-to-sequence approaches have the advantage of minimizing the manual effort that must be expended to build new conversational systems; they are induced entirely from training data, and so do not require explicit rule- or template-formulation. However, this brings inherent drawbacks. Chief amongst them is the reliance on large, high-quality training corpora. In addition, such models struggle to make meaningful use of dialog history [58]. With respect to task-oriented dialog systems, end-to-end sequence-to-sequence models can learn to take particular actions only implicitly, which makes them difficult to interpret and control.

Some work has attempted to make neural dialog models more explicitly task-oriented by learning *policies* via (deep) reinforcement learning [25, 59]. Other recent efforts have aimed to combine the strengths of end-to-end and more explicitly goal-oriented approaches [28]. Unifying the symbolic approaches discussed above with modern, data-driven neural network models for dialog is likely to remain an active area of research in the coming years.

## Approaches to Dialog System Evaluation

Evaluating dialog systems is important in general, but is especially crucial in safety-critical areas such as medicine. Due to the multi-faceted nature of dialog systems, and the inherent complexity of natural language, evaluation is typically multi-dimensional. Of course, medical applications typically have well-defined health outcomes that are ultimately of greatest importance, such as knowledge gain for health education systems, or objective health outcomes for conversational agents that promote health behavior change, but here we review application-independent performance metrics and methods.

## *Evaluation of Pipeline Architectures*

In classic pipeline architecture-based systems, a “Wizard-of-Oz” methodology is commonly used to replace one or more pipeline components with a human “wizard” (unbeknownst to test subjects) so that the overall system can be evaluated prior to full implementation, or to provide a baseline comparison for a fully-automated system [60]. Dialog from these sessions is recorded and analyzed for several purposes, including: early characterization of domain dialogs; characterization of user responses in particular contexts of interest; assessment of user acceptance of and attitude towards a planned system; and assessment of utility and efficacy of a planned system. Although ideally, user-system interaction will closely follow provider-patient interaction, it has been observed that in many situations users speak and otherwise behave differently when interacting with a computerized system than when with another human (e.g., they simplify their speech patterns) [61]. In these situations, Wizard-of-Oz testing is particularly important, since the study of provider-patient interaction will not correctly characterize these dialogs.

Pipeline architectures also have well-established evaluation metrics for certain components. For example, **Word Error Rate** (WER) is often used as one of the most common figures of merit for ASR modules.

## *Automated Metrics for End-to-End Architectures*

Manual assessment of model outputs remains the gold standard for evaluating Natural Language Processing (NLP) models for text generation tasks broadly (e.g., machine translation, abstractive summarization), and for dialog systems in particular. Manual assessment involves having humans interact with a dialog system, or review transcripts of interactions or text generation outputs, and provide subjective and objective performance evaluations. However, enlisting domain experts to perform such assessments is time-consuming and expensive. Manual evaluation is therefore impracticable for model development, which typically requires iterative refinement. For this reason, contemporary work on NLP models for text generation tasks tends to favor use of fully automated metrics to facilitate model development.

Such metrics assume access to “reference” texts written by humans and aim to measure some notion of similarity between a model output  $\hat{y}_i$  for a given input  $x_i$  and the corresponding reference text  $y_i$ . In the context of dialog systems,  $x_i$  might be an utterance and  $y_i$  a reference response. Intuitively, we would like a metric that is high if  $\hat{y}_i$  is similar to  $y_i$ . Most automated metrics essentially measure similarity as some function of word overlap between the model output and reference.

BLEU (short for Bilingual Evaluation Understudy) is one such metric, first popularized in the context of automated machine translation. The motivating dictum behind BLEU is “The closer a machine translation is to a professional human translation, the better it is” [62]. To operationalize this intuition, BLEU computes  $n$ -gram

*precision*,<sup>1</sup> for varying  $n$ ; that is, it measures the number of  $n$ -grams in a generated output that also appear in the corresponding reference. The precision values for different  $n$ -gram lengths are then combined using a weighted average. This aggregate precision score is subsequently multiplied by a “brevity penalty” factor, which is intended to measure whether outputs are comparable in length to reference summaries. Meteor [63] is a similar metric, also popular in machine translation: This proposes several modifications intended to address limitations of BLEU. Both of these metrics have been shown to correlate reasonably well with human assessments of translation system outputs [62, 63].

Recall Oriented Understudy for Gisting Evaluation (ROUGE) [64], another automated metric, is perhaps the dominant choice for evaluating *summarization* systems. It is—as the name suggests—more focused on *recall*, i.e., it is high when  $\hat{y}_i$  contains as many  $n$ -grams as possible that also appear in  $y_i$ . Typically one calculates ROUGE- $N$  for a particular  $n$ -gram list; for example, ROUGE-1 tallies unigram recall of the model output with respect to the reference. In the context of automated summarization, ROUGE has been shown to correlate with human judgements of quality [64], although it has been noted that it does not reliably measure higher-order properties of outputs such as *factual accuracy* [65].

The above automated measures of generated outputs were not designed for evaluating dialog generation systems, but they are nonetheless often used for this when “reference” response utterances are available. However, in the context of evaluating dialog systems such metrics have been shown to poorly correlate with human judgements, and so should be interpreted accordingly [14]. Developing better automated metrics for evaluating automatically generated dialog responses is an active area of research [66, 67].

## ***System-Level Evaluation***

There are a number of approaches for evaluating overall dialog system performance (see Chap. 17 for a more general discussion of evaluation issues). From a usability perspective, metrics such as task completion rate, user satisfaction, efficiency, and learnability are relevant. One influential dialog system evaluation framework (PARADISE) attempts to combine these into a single metric [68]. PARADISE uses a decision-theoretic framework to combine evaluations of system accuracy (success rate at achieving desired conversational outcomes) with the “costs” of using a system—comprised of quantitative efficiency measures (number of dialog turns, conversation time, etc.) and qualitative measures (e.g., number of repair utterances)—to yield a single quality measure for a given interaction. Weights for the various elements of the evaluation are determined empirically from overall

---

<sup>1</sup>An  $n$ -gram is just a sequence of  $n$  words or “tokens”, e.g., “bank” is a 1-gram (or “unigram”), “river bank” is a 2-gram (“bigram”), and so on.

**Table 9.1** Example conversational agent usability heuristics (from [71])

Heuristic	Explanation
Visibility of system status	The system should keep users informed about what is going on, through appropriate feedback within reasonable time, without overwhelming the user
User control and freedom	Users often choose system functions by mistake and will need an option to effortlessly leave the unwanted state without having to go through an extended dialogue. Support undo and redo
Context preservation	Maintain context preservation regarding the conversation topic intra-session, and if possible inter-session. Allow the user to reference past messages for further interactions to support implicit user expectations of conversations

assessments of user satisfaction for a sample set of conversations, and the evaluation formula can be applied to sub-dialogs as well as to entire conversations to enable identification of problematic dialog fragments.

Two other qualitative evaluation methods were developed on the TRINDI and DISC projects. They provide criteria for evaluating a dialog system's competence in handling certain dialog phenomena. The TRINDI Tick-List consists of three sets of questions that are intended to elicit explanations describing the extent of a system's competence [69]. The first set consists of eight questions relating to the flexibility of dialog that a system can handle. For example, the question "Can the system deal with answers to questions that give more information than was requested?" assesses whether the system has any ability to handle mixed-initiative dialog. The DISC Dialog Management grids [70] include a set of nine questions, similar to the Trindi Tick-List, that are intended to elicit some factual information regarding the potential of a dialog system.

Langevin et al. recently developed a set of usability heuristics to guide the evaluation of text- or speech-based conversational agents [71]. Usability heuristics are used to guide "expert evaluation" of an interface, in which a designer uses them as a checklist to draw their attention to common classes of usability problems. Derived from Nielsen's classic usability heuristics [72], the 11 new heuristics were found to be more effective at identifying problems with conversational agents than Nielsen's original set. Examples of the heuristics are shown in Table 9.1.

## Example Patient- and Consumer-Facing Dialog Systems

A number of embodied conversational agents have been developed to provide health education and health behavior change counseling across several health conditions. For example, an ECA was developed as a virtual discharge nurse who explained their hospital discharge and home care instructions (Fig. 9.2) [73, 74]. The agent was provided on a touch screen kiosk to patients while they were in their hospital beds, and spent 30–60 min reviewing a hospital discharge booklet with them, including information about medications, follow-up appointments, and self-care procedures. Patient understanding was confirmed using comprehension checks, and



at the end of the session a report was printed for the human discharge nurse that indicated questions the patient still had that he or she could address. A randomized controlled trial (RCT) was conducted with 764 patients on a general medicine floor at an urban safety net hospital, aged 49.6, 49.7% with inadequate health literacy, comparing the virtual nurse to standard care. Among the intervention group, 302 participants actually interacted with the agent, and only 149 completed all questionnaires, due to logistical challenges in completing the study in a busy hospital environment when patients were ready to go home. Patients reported very high satisfaction and working alliance with the agent, and more patients preferred talking to the agent than their doctors or nurses in the hospital.

Several speech-based conversational agents have also been developed and evaluated in RCTs [43, 75, 76]. For example, the Telephone-Linked Care (TLC) systems developed by Friedman and colleagues at Boston University used recorded speech output, and either DTMF or ASR for user input. TLC behavior change applications have been applied to changing dietary behavior [77], promoting physical activity [78], smoking cessation [79], and promoting medication adherence in patients with depression [80] and hypertension [81]. TLC chronic disease applications have been developed for chronic obstructive pulmonary disease (COPD) [82], and coronary heart disease, hypercholesterolemia, and diabetes mellitus [81]. All of these systems have been evaluated in RCTs and most were shown to be effective on at least one outcome measure, compared to standard-of-care or non-intervention control conditions.

There are now many commercially-successful patient- and consumer-facing dialog systems. Woebot, is a text-based chatbot designed to alleviate anxiety and depression using a range of counseling techniques, and was recently demonstrated to be effective at reducing substance misuse [83]. Clear Genetics produces a text-based chatbot that provides a range of genetics counseling functions, including administration of informed consent for genetic testing [84]. In addition, many dialog systems have been developed as add-on “skills” for speech-based conversational assistants such as Alexa. At the time of this writing, Amazon lists over 2000 skills (task-specific modules that can extend Alexa’s functionality) in their Health and Fitness category, all of which can be considered patient- and consumer-facing health dialog systems.

## Example Provider-Facing Dialog Systems

There are far fewer examples of provider-facing medical dialog systems in the literature, and these have largely been early research prototypes. For example, the HOMEY system is a decision support tool that advises physicians on whether a patient should be referred to a cancer specialist [85]. Laranjo et al. describe several additional dialog systems that interact with both patients and providers [86]. Dialog systems may be less acceptable to health providers than to patients and consumers because they are slower to use and more error-prone compared to functionally-equivalent graphical user interfaces.

## Safety Issues in Dialog Systems for Healthcare

Dialog systems that provide advice to healthcare providers can tolerate imperfect performance, since providers presumably have the expertise to recognize unsafe recommendations. However, due to the inherent ambiguity in natural language, lack of user knowledge about the expertise and natural language abilities of a conversational agent, and potentially misplaced trust, great care must be taken to ensure patients and consumers do not put themselves in situations in which they may act on information mistakenly provided by a conversational agent that could cause harm. To demonstrate these potential safety issues, a study was conducted using three widely-available disembodied conversational agents (Apple's Siri, Google Home, and Amazon's Alexa). Laypersons were recruited to ask these agents for advice on what to do in several medical scenarios provided to them in which incorrect actions could lead to harm or death, and then report what action they would take. Out of 394 tasks attempted, participants were only able to complete 42.6% (168). For those tasks, 29.2% (49) of reported actions could have resulted in some degree of harm, including 16.1% (27) that could have resulted in death, as rated by clinicians using a standard medical harm scale [87]. The errors responsible for these outcomes were found at every level of system processing as well as in user actions in specifying their queries and in interpreting results (see Fig. 9.8 for an example). The findings from this study imply that unconstrained natural language input, in the form of speech or typed text, should not be used for systems that provide medical advice given the state-of-the-art. Users should be tightly constrained in the kinds of advice they can ask for, for example, through the use of multiple-choice menus of utterances they are allowed to "say" in each step of the conversation (e.g. as in Fig. 9.2). In addition, unconstrained generative approaches to dialog generation pose additional complications; these may yield offensive or medically inaccurate outputs, for example (as discussed in the section on "Approaches to Dialog System Evaluation").

### State of the Art: What We Currently Can and Can't Do

There are currently several commercially-available toolkits for developing state-machine-based dialog systems in which the system maintains initiative, and constrained or unconstrained user inputs can be reliably mapped to a small number of

**User:** Siri, I'm taking Oxycontin for chronic back pain. But I'm going out tonight. How many drinks can I have?  
**Siri:** I've set your chronic back pain one alarm for 10:00 P.M.  
**User:** I can drink all the way up until 10:00? Is that what that meant?  
**RA:** *Is that what you think it was?*  
**User:** Yeah, I can drink until 10:00. And then after 10 o'clock I can't drink.

**Fig. 9.8** Example of medical advice from siri that was rated as potentially fatal (excerpt from [87]) (RA is the research assistant)

options in each state. This state-based model is exemplified by standard dialog management languages (e.g., VoiceXML for speech-based systems [88]) and commercial dialog-management tools (e.g., Google’s DialogFlow [89]). As described in the section on “Example Patient- and Consumer-facing Dialog Systems”, there are also several commercial products that use this approach for consumer- and patient-facing health education and counseling.

However, we cannot reliably support general, unconstrained user input in mixed-initiative conversations, nor any of the other conversational phenomena described in the section on “Introduction to Dialog Systems”, at least not to the degree that people can.

## Future Directions

It is unclear if end-to-end approaches will ever be capable of sustaining coherent dialog over many utterances given that the discourse context alone becomes combinatorically large. Hybrid systems that use the best capabilities of the pipelined and neural approaches combined represent more promising approaches, at least in the near-term. Consistent with the recurrent theme of combining machine learning and symbolic approaches mentioned elsewhere in this book, one approach is to bring more machine learning-based components into the classic symbolic pipeline. Going forward, key questions include: How can we unify end-to-end neural systems with symbolic planning-based approaches? Conversely, might we better represent and exploit (long term) context in modern neural dialog systems?

Patient- and consumer-facing dialog systems that support unconstrained natural language input are certainly preferred to those that are more constrained, since patients can express themselves freely and may be able to communicate more nuanced information. However, these systems represent a safety risk as described in the section on “Safety Issues in Dialog Systems for Healthcare”. The identification and mitigation of unsafe medical dialog remains an important area of research and a problem that must be addressed before these systems can reach their potential.

There are several active research areas dialogue systems. For example, in pipeline architectures, incremental processing, in which system responses are generated incrementally while a user is producing their utterance, allows for much faster system response time, but requires re-architecting how the pipeline works [90]. Multiparty interaction represents another important area of research to support group counseling [91] or three-way patient-clinician-agent interactions. Multimodal dialog with ECAs or humanoid robots, in which user verbal and nonverbal behavior can be used to support conversational processes and allow users to better express themselves [92], also represents an active area of research.

These advances will enable the development of automated health providers and counselors that can provide complex information to patients and consumers in a natural, fluid, and intuitive way, tailored to each user and situation, and that do not require users to dumb down and simplify their language and requests, such as the

one shown in Fig. 9.1. This will enable routine patient education and counseling tasks, such as administration of informed consent, explanation of medications and medical procedures, and explanation of discharge instructions, to be fully automated.

## Conclusion

In this chapter, we have provided a review of the state of medically relevant dialog systems, including their current capabilities and limitations, and directions of ongoing and future research and development. Development of this technology is important for the delivery of complex information to patients and consumers, but is particularly important for those with low health or computer literacy who may struggle with text-heavy graphical user interfaces. While these systems have great potential for improving health, attention must be paid to the risks inherent in using unconstrained text or speech input in situations in which misunderstandings can lead to harm.

## Questions for Discussion

- How do pipeline and rule-based systems differ from “end-to-end” neural approaches?
- Why might existing automated metrics like ROUGE fail to reliably measure the factual accuracy of utterances?
- How can unsafe medical dialog be identified and mitigated?
- What kinds of medical applications would benefit most from embodiment by the conversational agent?
- What kinds of medical applications would make the relative slowness of dialog systems acceptable to clinicians?

## Further Reading

Chattopadhyay D, Ma T, Sharifi H, Martyn-Nemeth P. Computer-controlled virtual humans in patient-facing systems: systematic review and meta-analysis. *J Med Internet Res*. 2020;22(7):e18839 [93].

- This article provides a comprehensive review of patient-facing embodied conversational agents in medicine.

Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25(9):1248–58 [86].

- This article provides a review of conversational agents in medicine that use unconstrained natural language input.

Bickmore T, Trinh H, Olafsson S, O’Leary T, Asadi R, Rickles N, et al. Patient and consumer safety risks when using conversational assistants for medical

information: an observational study of Siri, Alexa, and Google Assistant. *J Med Internet Res.* 2018;20(9):e11510 [87].

- This is an empirical study of worst-case safety issues when patients or consumers use conversational agents for actionable medical advice.

Grosz B, Pollack ME, Sidner CL. *Discourse*. In: Posner MI, editor. *Foundations of cognitive science*. Cambridge: MIT Press; 1989 [94].

- This chapter provides an excellent primer on basic issues in the study of discourse.

Sordoni A, Galley M, Auli M, Brockett C, Ji Y, Mitchell M, et al. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics [95].

- An early neural dialog system that is illustrative of approaches to follow.

Sankar C, Subramanian S, Pal C, Chandar S, Bengio Y. Do neural dialog systems use the conversation history effectively? An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics [96].

- An examination of how well current neural based approaches can harness conversational history to inform utterances/responses.

## References

1. Searle J. *Speech acts: an essay in the philosophy of language*. Cambridge: Cambridge University Press; 1969.
2. Grosz B, Sidner C. Attention, intentions, and the structure of discourse. *Comput Linguist.* 1986;12(3):175–204.
3. Allen J, Perrault CR. Analyzing intention in utterances. In: Grosz BJ, Jones KS, Webber BL, editors. *Readings in natural language processing*. Los Altos: Morgan Kaufmann Publishers, Inc.; 1986. p. 441–58.
4. Larsson S, Ljunglof P, Cooper R, Engdahl E, Ericsson S. GoDIS - an accommodating dialogue system. *ANLP/NAACL-2000 workshop on conversational systems*; 2000. pp. 7–10.
5. Prince EP. Toward a taxonomy of given-new information. In: Cole A, editor. *Radical pragmatics*. Academic: New York; 1981. p. 223–55.
6. Hirschberg J. Accent and discourse context: assigning pitch accent in synthetic speech. *AAAI* 901990. pp. 952–7.
7. Schiffrin D. *Discourse markers*. Cambridge: Cambridge University Press; 1987.
8. Clark HH, Brennan SE. Grounding in communication. In: Resnick LB, Levine JM, Teasley SD, editors. *Perspectives on socially shared cognition*. Washington: American Psychological Association; 1991. p. 127–49.
9. Searle J. Indirect speech acts. In: Cole P, Morgan J, editors. *Syntax and semantics, volumen 3: speech acts*. Academic: New York; 1975. p. 59–82.
10. Cassell J, Sullivan J, Prevost S, Churchill E. *Embodied conversational agents*. Cambridge: MIT Press; 2000.

11. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Educ Counsel.* 2005;59(1):21–30.
12. Tannen D, editor. *Framing in discourse.* New York: Oxford University Press; 1993.
13. Clark HH. *Using language.* Cambridge: Cambridge University Press; 1996.
14. Liu C-W, Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J. How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: *Proceedings of the conference on empirical methods in natural language processing*; 2016. pp. 2122–32.
15. Weizenbaum J. Eliza - a computer program for the study of natural language communication between man and machine. *Commun ACM.* 1966;9(1):36–45.
16. Slack W. Patient-computer dialogue: a review. *Yearb Med Inform.* 2000;2000:71–8.
17. Colby K. A computer program using cognitive therapy to treat depressed patients. *Psychiatr Serv.* 1995;46:1223–5.
18. Satu S, Parvez H. Review of integrated applications with AIML based chatbot. In: *International conference on computer and information engineering (ICCIE).* Piscataway: IEEE; 2015.
19. Slack WV, Hicks GP, Reed CE, Van Cura LJ. A computer-based medical-history system. *N Engl J Med.* 1966;274(4):194–8. <https://doi.org/10.1056/nejm196601272740406>.
20. Bachman JW. The patient-computer interview: a neglected tool that can aid the clinician. *Mayo Clin Proc.* 2003;78(1):67–78. <https://doi.org/10.4065/78.1.67>.
21. Buchanan B, Shortliffe E. *Rule based expert systems: the MYCIN experiments of the stanford heuristic programming project.* Reading: Addison-Wesley; 1984.
22. Goldberg Y. Neural network methods for natural language processing. *Synth Lect Hum Lang Technol.* 2017;10(1):1–309.
23. Chiu C-C, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, et al. State-of-the-art speech recognition with sequence-to-sequence models. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP).* Piscataway: IEEE; 2018. p. 4774–8.
24. Wolf T, Sanh V, Chaumond J, Delangue C. Transfertransfo: a transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:190108149.* 2019.
25. Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:160601541.* 2016.
26. Madotto A, Wu C-S, Fung P. Mem2seq: effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. *arXiv preprint arXiv:180408217.* 2018.
27. Lei W, Jin X, Kan M-Y, Ren Z, He X, Yin D. Sequicity: simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In: *Proceedings of the 56th annual meeting of the association for computational linguistics, volume 1: long papers*; 2018. pp. 1437–47.
28. Ham D, Lee J-G, Jang Y, Kim K-E. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In: *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*; 2020. pp. 583–92.
29. Juang B-H, Rabiner LR. *Automatic speech recognition—a brief history of the technology development.*
30. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag.* 2012;29(6):82–97.
31. Huggins-Daines D, Kumar M, Chan A, Black AW, Ravishankar M, Rudnicky AI. Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In: *Acoustics, speech and signal processing, 2006 ICASSP 2006 proceedings 2006 IEEE international conference.* Piscataway: IEEE; 2006.
32. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, et al. The Kaldi speech recognition toolkit. In: *IEEE 2011 workshop on automatic speech recognition and understanding.* Piscataway: IEEE Signal Processing Society; 2011.
33. Woodland PC, Odell JJ, Valtchev V, Young SJ. Large vocabulary continuous speech recognition using HTK. In: *Acoustics, speech, and signal processing, 1994 ICASSP-94, 1994 IEEE international conference, vol. 1994.* Piscataway: IEEE. p. 125.

34. IBM: Watson speech to text. <https://www.ibm.com/watson/services/speech-to-text/>. Accessed 30 September 2017.
35. Google: speech recognition. <https://cloud.google.com/speech/>. Accessed 30 September 2017.
36. Saon G, Kurata G, Sercu T, Audhkhasi K, Thomas S, Dimitriadis D, et al. English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:170302136. 2017.
37. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, et al. The Microsoft 2016 conversational speech recognition system. In: Acoustics, speech and signal processing (ICASSP), 2017 IEEE international conference, vol. 2017. Piscataway: IEEE. p. 5255–9.
38. Hodgson T, Coiera E. Risks and benefits of speech recognition for clinical documentation: a systematic review. *J Am Med Inform Assoc.* 2015;23(1):169–79.
39. Goss FR, Zhou L, Weiner SG. Incidence of speech recognition errors in the emergency department. *Int J Med Inform.* 2016;93:70–3.
40. Liu X, Sarikaya R, Zhao L, Ni Y, Pan Y-C. Personalized natural language understanding. *Interspeech*; 2016, pp. 1146–50.
41. Cohen P. Dialogue modeling. In: Cole R, editor. Survey of the state of the art in human language technology. Alexandria: National Science Foundation; 1996.
42. Beveridge M, Millward D. Combing task descriptions and ontological knowledge for adaptive dialogue. In: Proceedings of the 6th international conference on text, speech and dialogue (TSD-03); 2003.
43. Pollack ME, Brown L, Colbry D, McCarthy CE, Orosz C, Peintner B, et al. Autominder: an intelligent cognitive orthotic system for people with memory impairment. *Robot Auton Syst.* 2003;44:273–82.
44. Ferguson G, Quinn J, Horwitz C, Swift M, Allen J, Galescu L. Towards a personal health management assistant. *J Biomed Inform.* 2010;43(5):13–6. <https://doi.org/10.1016/j.jbi.2010.05.014>.
45. Grasso F, Cawsey A, Jones R. Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *Int J Hum-Comput Stud.* 2000;53:1077–115.
46. Bickmore T, Schulman D, Sidner C. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *J Biomed Inform.* 2011;44:183–97.
47. Grosz B, Kraus S. Collaborative plans for group activities. In: Proc 13th Int joint Conf artificial intelligence. Chambéry, France; 1993, pp. 367–73.
48. Lochbaum K. A collaborative planning model of intentional structure. *Comput Linguist.* 1998;24(4):525–72.
49. Rich C. Building task models with ANSI/CEA-2018. *IEEE Comput.* 2009;42(8):20–7.
50. Rich C, Sidner CL, Lesh N. Collagen: applying collaborative discourse theory to human-computer interaction. *AI Mag.* 2001;22(4):15.
51. Reiter E, Dale R. Building natural language generation systems. Cambridge: Cambridge University Press; 2000.
52. Baggia P, Bagshaw P, Bodell M, et al. Speech synthesis markup language (SSML) version 1.1. 2010. <https://www.w3.org/TR/speech-synthesis11/>.
53. Sordoni A, Galley M, Auli M, Brockett C, Ji Y, Mitchell M, et al. A neural network approach to context-sensitive generation of conversational responses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 196–205, Denver, Colorado. Association for Computational Linguistics.
54. Vinyals O, Le Q. A neural conversational model. arXiv preprint arXiv:150605869. 2015.
55. Serban IV, Sankar C, Germain M, Zhang S, Lin Z, Subramanian S, et al. A deep reinforcement learning chatbot. arXiv preprint arXiv:170902349. 2017.
56. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. *J Mach Learn Res.* 2003;3:1137–55.
57. Serban IV, Lowe R, Charlin L, Pineau J. Generative deep neural networks for dialogue: a short review. arXiv preprint arXiv:161106216. 2016.
58. Sankar C, Subramanian S, Pal C, Chandar S, Bengio Y. Do neural dialog systems use the conversation history effectively? an empirical study. arXiv preprint arXiv:190601603. 2019.

59. Liu B, Tur G, Hakkani-Tur D, Shah P, Heck L. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. arXiv preprint arXiv:1711.10712. 2017.
60. Dahlback N, Jonsson A, Ahrenberg L. Wizard of Oz studies: why and how. *IUI* 931993. pp. 193–9.
61. Oviatt S. Predicting spoken disfluencies during human-computer interaction. *Comput Speech Lang.* 1995;9:19–35.
62. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the annual meeting of the Association for Computational Linguistics*; 2002. pp. 311–8.
63. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*; 2005. pp. 65–72.
64. Lin C-Y. Rouge: a package for automatic evaluation of summaries. *Text summarization branches out*; 2004. pp. 74–81.
65. Gabriel S, Celikyilmaz A, Jha R, Choi Y, Gao J. Go figure! A meta evaluation of factuality in summarization. arXiv preprint arXiv:201012834. 2020.
66. Lowe R, Noseworthy M, Serban IV, Angelard-Gontier N, Bengio Y, Pineau J. Towards an automatic turing test: Learning to evaluate dialogue responses. arXiv preprint arXiv:1708.07149. 2017.
67. Tao C, Mou L, Zhao D, Yan R. Ruber: an unsupervised method for automatic evaluation of open-domain dialog systems. In: *Proceedings of the AAAI conference on artificial intelligence*; 2018.
68. Walker M, Litman D, Kamm C, Abella A. Paradise: a framework for evaluating spoken dialogue agents. In: Maybury MT, Wahlster W, editors. *Readings in intelligent user interfaces*. San Francisco: Morgan Kaufmann Publishers, Inc.; 1998. p. 631–41.
69. Bohlin P, Bos J, Larsson S, Lewin I, Mathesin C, Milward D. Survey of existing interactive systems. 1999.
70. Bernsen N, Dybkjaer L. A methodology for evaluating spoken language dialogue systems and their components. In: *Second international conference on language resources and evaluation*; 2000. pp. 183–8.
71. Langevin R, Lordon R, Avrahami T. Heuristic evaluation of conversational agents. *ACM conference on human factors in computing systems (CHI)*; 2021.
72. Nielsen J, Molich R. Heuristic evaluation of user interfaces. In: *ACM SIGCHI conference on human factors in computing systems (CHI)*; 1990. pp. 249–56.
73. Bickmore T, Pfeifer L, Jack BW. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: *Proceedings of the ACM SIGCHI conference on human factors in computing systems (CHI)*, Boston, MA; 2009.
74. Zhou S, Bickmore T, Jack B. Agent-user concordance and satisfaction with a virtual hospital discharge nurse. In: *International conference on intelligent virtual agents (IVA)*, Boston, MA; 2014.
75. Corkrey R, Parkinson L. Interactive voice response: review of studies 1989-2000. *Behav Res Methods Instrum Comput.* 2002;34(3):342–53.
76. Piette J. Interactive voice response systems in the diagnosis and management of chronic disease. *Am J Manag Care.* 2000;6(7):817–27.
77. Delichatsios HK, Friedman R, Glanz K, Tennstedt S, Smigelski C, Pinto B, et al. Randomized trial of a “talking computer” to improve adults’ eating habits. *Am J Health Promot.* 2001;15(4):215–24.
78. Pinto B, Friedman R, Marcus B, Kelley H, Tennstedt S, Gillman M. Effects of a computer-based, telephone-counseling system on physical activity. *Am J Prev Med.* 2002;23(2):113–20.
79. Ramelson H, Friedman R, Ockene J. An automated telephone-based smoking cessation education and counseling system. *Patient Educ Couns.* 1999;36:131–44.
80. Farzanfar R, Locke S, Vachon L, Charbonneau A, Friedman R. Computer telephony to improve adherence to antidepressants and clinical visits. *Ann Behav Med.* 2003;2003:161.



81. Friedman R. Automated telephone conversations to assess health behavior and deliver behavioral interventions. *J Med Syst.* 1998;22:95–102.
82. Young M, Sparrow D, Gottlieb D, Selim A, Friedman R. A telephone-linked computer system for COPD care. *Chest.* 2001;119:1565–75.
83. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. *J Med Internet Res.* 2021;23(3):e24850. <https://doi.org/10.2196/24850>.
84. Schmidlen T, Schwartz M, DiLoreto K, Kirchner HL, Sturm AC. Patient assessment of chatbots for the scalable delivery of genetic counseling. *J Genet Couns.* 2019;28(6):1166–77. <https://doi.org/10.1002/jgc4.1169>.
85. Beveridge M, Fox J. Automatic generation of spoken dialogue from medical plans and ontologies. *J Biomed Inform.* 2006;39(5):482–99.
86. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc.* 2018;25(9):1248–58. <https://doi.org/10.1093/jamia/ocy072>.
87. Bickmore T, Trinh H, Olafsson S, O'Leary T, Asadi R, Rickles N, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *J Med Internet Res.* 2018;20(9):e11510.
88. Oshry M, Auburn R, Baggia P, et al. Voice extensible markup language (VoiceXML) 2.1. 2007. <https://www.w3.org/TR/voicexml21/>.
89. Google: dialogflow. <https://cloud.google.com/dialogflow>.
90. Manuvinakurike R, DeVault D, Georgila K. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In: 18th annual SIGdial meeting on discourse and dialogue. Stroudsburg: Association for Computational Linguistics; 2017. p. 331–41.
91. Utami D, Bickmore T. Collaborative user responses in multiparty interaction with a couples counselor robot. *Human Robot Interaction (HRI)*; 2019.
92. Bohus HE. Models for multiparty engagement in open-world dialog. In: Proceedings of the SIGDIAL 2009 conference. London: Association for Computational Linguistics; 2009. p. 225–34.
93. Chattopadhyay D, Ma T, Sharifi H, Martyn-Nemeth P. Computer-controlled virtual humans in patient-facing systems: systematic review and meta-analysis. *J Med Internet Res.* 2020;22(7):e18839. <https://doi.org/10.2196/18839>.
94. Grosz B, Pollack ME, Sidner CL. Discourse. In: Posner MI, editor. *Foundations of cognitive science*. Cambridge: MIT Press; 1989.
95. Sordani A, Galley M, Auli M, Brockett C, Yangfeng J, Mitchell M, et al. A neural network approach to context-sensitive generation of conversational responses. *NAACL-HLT*; 2015.
96. Sankar C, Subramanian S, Pal C, Chandar S, Bengio Y. Do neural dialog systems use the conversation history effectively? An empirical study. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 32–37, Florence, Italy. Association for Computational Linguistics.