# Survey on evaluation methods for dialogue systems

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, Mark Cieliebak

Artificial Intelligence Review (2021) 54:755–810
https://doi.org/10.1007/s10462-020-09866-x

# Survey on evaluation methods for dialogue systems

**Jan Deriu[1] · Alvaro Rodrigo[2] · Arantxa Otegi[3] · Guillermo Echegoyen[2] · Sophie Rosset[4] · Eneko Agirre[3] · Mark Cieliebak[1]**

Published online: 25 June 2020


**Abstract**
In this paper, we survey the methods and concepts developed for the evaluation of dialogue systems. Evaluation, in and of itself, is a crucial part during the development process. Often, dialogue systems are evaluated by means of human evaluations and questionnaires. However, this tends to be very cost- and time-intensive. Thus, much work has been put into finding methods which allow a reduction in involvement of human labour. In this survey, we present the main concepts and methods. For this, we differentiate between the various classes of dialogue systems (task-oriented, conversational, and question-answering dialogue systems). We cover each class by introducing the main technologies developed for the dialogue systems and then present the evaluation methods regarding that class.

**Keywords** Dialogue systems · Evaluation metrics · Discourse model · Conversational AI · Chatbots

✉ Jan Deriu
deri@zhaw.ch

Alvaro Rodrigo
alvarory@lsi.uned.es

Arantxa Otegi
arantza.otegi@ehu.eus

Guillermo Echegoyen
gblanco@lsi.uned.es

Sophie Rosset
sophie.rosset@limsi.fr

Eneko Agirre
e.agirre@ehu.eus

Mark Cieliebak
ciel@zhaw.ch

[1] Zurich University of Applied Sciences (ZHAW), Steinberggasse 13, 8400 Winterthur, Switzerland

[2] NLP & IRGroup, UNED, C/Juan del Rosal 16, 28040 Madrid, Spain

[3] IXA NLP Group, University of the Basque Country(UPV/EHU), Manuel Lardizabal 1 Donostia, 20018 Basque Country, Spain

[4] CNRS, LIMSI, Université Paris-Saclay, Campus Universitaire, Bât. 508, rue John von Neumann, 91405 Orsay Cedex, France

## 1 Introduction

As the amount of digital data continuously grows, users demand technologies that offer quick access to such data. In fact, users rely on systems that support information search interactions such as Siri[1], Google Assistant[2], Amazon Alexa[3] or Microsoft XiaoIce (Zhou et al. 2018), etc. These technologies, called Dialogue Systems (DS), allow the user to converse with a computer system using natural language. Dialogue Systems are applied to a variety of tasks, e.g.:

- Virtual Assistants aid users in everyday tasks, such as scheduling appointments. They usually operate on predefined actions which can be triggered by voice command.
- Information-seeking systems provide users with information about a question (e.g. the most suitable hotel in town). These questions also include factual questions as well as more complex questions.
- E-learning dialogue systems train students for various situations. For instance, they train the interaction with medical patients or train military personnel in questioning a witness.

One crucial step in the development of DS is evaluation. That is, to measure how well the DS is performing. However, evaluating a dialogue system can prove to be problematic because there are two important factors to be considered. Firstly, the definition of what constitutes a high-quality dialogue is not always clear and often depends on the application. Even if a definition is assumed, it is not always clear how to measure it. For instance, if we assume that a high-quality dialogue system is defined by its ability to respond with an appropriate utterance, it is not clear how to measure appropriateness or what appropriateness means for a particular system. Moreover, one might ask the users if the responses were appropriate, but as we will discuss below, user feedback might not always be reliable for a variety of reasons.

The second factor is that the evaluation of dialogue systems is very cost- and time-intensive. This is especially true when the evaluation is carried out by a user study, which requires careful preparation, the need for inviting and compensating users for their participation.

Over the past decades, many different evaluation methods have been proposed. The evaluation methods are closely tied to the characteristics of the dialogue system which they are aimed at evaluating. Thus, quality is defined in the context of the function which dialogue system is meant to fulfil. For instance, a system designed to answer questions will be evaluated on the basis of correctness, which is not necessarily a suitable metric for evaluating a conversational agent.

Most methods are aimed at automating the evaluation, or at least automating certain aspects of the evaluation. The goal of an evaluation method is to obtain automated and repeatable evaluation procedures that allow efficient comparisons in the quality of different dialogue strategies.

[1] https://www.apple.com/es/siri/.
[2] https://assistant.google.com/.
[3] https://www.amazon.com.

**Table 1** Characterizations of the different dialogue system types

| | Task-oriented DS | Conversational agents | Interactive QA |
|---|---|---|---|
| Task | Yes—clear defined | No | Yes—answer questions |
| Dial. Structure | Highly structured | Not structured | No |
| Domain | Restricted | Mostly open domain | Mixed |
| Turns | Multi | Multi | Single/Multi |
| Length | Short | Long | – |
| Initiative | Mixed/system init | Mixed/user init | User init |
| Interface | Multi-modal | Multi-modal | Mostly text |

This survey is structured as follows; in the next section we give a general overview over the different classes of dialogue systems and their characteristics. We then introduce the evaluation task in greater detail, with an emphasis on the goals of an evaluation and the requirements on an evaluation metric. In Sects. 3, 4, and 5, we introduce each dialogue system class (i.e. task-oriented systems, conversational agents and question answering dialogue systems). Thereafter, we give an overview of the characteristics, dialogue behaviour, and concepts behind the implementation methods of the various dialogue systems. Finally, we present the evaluation methods and the ideas behind them. Here, we set an emphasis the relationship between these methods and the dialogue system classes, including which aspects of the evaluation are automated. In Sect. 6, we give a short overview of the relevant datasets and evaluation campaigns in the domain of dialogue systems. In Sect. 7, we discuss the issues and challenges in devising automated evaluation methods and discuss the level of automation achieved.

## 2 A general overview

### 2.1 Dialogue systems

Dialogue Systems (DS) usually structure dialogues in *turns*, each turn is defined by one or more *utterances* from one speaker. Two consecutive turns between two different speakers is called an *exchange*. Multiple exchanges constitute a *dialogue*. Another different, but related view is to interpret each turn or each utterance as an action (more on this later).

The main component of a dialogue system is the dialogue manager that defines the content of the next utterance and thus the behaviour of the dialogue system. There are many different approaches to design a dialogue manager, which are partly dictated by the application of the dialogue system. However, there are three broad classes of dialogue systems that we encounter in the literature: task-oriented systems, conversational agents and interactive question answering systems[4].

[4] In recent literature, the distinction is made only between the first two classes of dialogue systems (Serban et al. 2018; Chen et al. 2017; Jurafsky and Martin 2017). However, interactive question answering systems cannot be completely placed in either of the two categories.

We identified the following characteristic features that help differentiate between the three different classes: whether the system is developed to solve a task, whether the dialogue follows a structure, whether the domain is restricted or open, whether the dialogue spans over multiple turns, whether the dialogues are long or rather efficient, who takes the initiative, and what interface is used (text, speech, multi-modal). Table 1 depicts the characteristics for each of the dialogue system classes. In this table, we can see the following main features for each class:

- Task-oriented systems are developed to help the user solve a specific task as efficiently as possible. The dialogues are characterized by following a clearly defined structure that is derived from the domain. The dialogues follow mixed initiative; both the user and the system can take the lead. Usually, the systems found in the literature are built for speech input and output. However, task-oriented systems in the domain of assisting users are built on multi-modal input and output.
- Conversational agents display a more unstructured conversation, as their purpose is to have open-domain dialogues with no specific task to solve. Most of these systems are built to emulate social interactions, and thus longer dialogues are desired.
- Question Answering (QA) systems are built for the specific task of answering questions. The dialogues are not defined by a structure as with task-oriented systems, however, they mostly follow the question and answer style pattern. QA systems may be built for a specific domain, but may be also tilted towards more open domain questions. Usually, the domain is dictated by the underlying data, e.g. knowledge bases or text snippets from forums. Traditional QA systems work on a singe-turn interaction, however, there are systems that allow multiple turns to cover follow-up questions. The initiative is mostly done by the user, who asks questions.

## 2.2 Evaluation

Evaluating dialogue systems is a challenging task and subject of much research. We define the goal of an evaluation method as having an automated, repeatable evaluation procedure with high correlation to human judgments, which is able to differentiate between various dialogue strategies and is able to explain which features of the dialogue systems are important. Thus, the following requirements can be stated:

- *Automatic* in order to reduce the dependency on human labour, which is time- and cost-intensive as well as not necessarily repeatable, the evaluation method should be automated, or at least partially automated.
- *Repeatable* the evaluation method should yield the same result if applied multiple times to the same dialogue system under the same circumstances.
- *Correlated to human judgments* the procedure should yield ratings that correlate to human judgments.
- *Differentiate between different dialogue systems* the evaluation procedure should be able to differentiate between different strategies. For instance, if one wants to test the effect of a *barge-in* feature (i.e. allowing the user to interrupt the dialogue system), the evaluation procedure should be able to highlight the effects.

- *Explainable* the method should give insights into which features of the dialogue system impact the quality of the dialogue and in which manner they do so. For instance, the methods should reveal that the automatic speech recognition system's *word-error rate* has a high influence on the quality of the natural language understanding component, which in turn impacts the intent classification.

In this survey, we focus on the efforts of automating the evaluation process. This is a very difficult, but crucial task, as human evaluations are cost- and time-intensive. Although much progress has been made in automating the evaluations of dialogue systems, the reliance on human evaluation is still present. Here, we give a condensed overview on the human-based evaluations used in the literature.

*Human evaluation* There are various approaches to a human evaluation. The test subjects can take on two main roles: interacting with the system or rating a dialogue or utterance, or both. In the following, we differentiate among different types of user populations. Among each of the populations, the subjects can take on any of the two roles.

- *Lab experiments* Before crowdsourcing was popular, dialogue systems were evaluated in a lab environment. Users were invited to participate in the lab where they interacted with a dialogue system and subsequently filled a questionnaire. For instance, Young et al. (2010) recruited 36 subjects, which were given instructions and presented with various scenarios. The subjects were asked to solve a task using a spoken dialogue system. Furthermore, a supervisor was present to guide the users. The lab environment is very controlled, which is not necessarily comparable to the real world (Black et al. 2011; Schmitt and Ultes 2015).
- *In-field experiments* Here, the evaluation is performed by collecting feedback from real users of the dialogue systems (Lamel et al. 2000). For instance, for the Spoken Dialogue Challenge (Black et al. 2011), the systems were developed to provide bus schedule information in Pittsburgh. The evaluation was performed by redirecting the evening calls to the dialogue systems and getting the user feedback at the end of the conversation. The Alexa Prize[5] also followed the same strategy, i.e. it let real users interact with operational systems and gathered user feedback over a span of several months.
- *Crowdsourcing* Recently, human evaluation has shifted from a lab environment to using crowdsourcing platforms such as Amazon Mechanical Turk (AMT). These platforms provide large amounts of recruited users. Jurcícek et al. (2011) evaluate the validity of using crowdsourcing for evaluating dialogue systems, and their experiments suggest that using enough crowdsourced users, the quality of the evaluation is comparable to the lab conditions. Current research relies on crowdsourcing for human evaluation (Serban et al. 2017a; Wen et al. 2017).

  Especially conversational dialogue systems are evaluated via crowdsourcing, where there are two main evaluation procedures: crowdworkers either talk to the system and rate the interaction or they are presented with a context from the test set and a response by the system, which they need to rate. In both settings, the crowdworkers are aksed to rate the system based on quality, fluency or appropriateness. Recently, Adiwardana et al. (2020) introduced Sensibleness and Specificity Average (SSA), where humans rate the sensibleness and specificity of a response. These capture two aspects of human

[5] https://developer.amazon.com/alexaprize.

behaviour: making sense and being specific. A dialogue system can be sensible by responding with vague answers (e.g. "I don't know"), whereas it is only specific if it takes the context into account.

Human based evaluation is difficult to set up and to carry out. Much care has to be taken in setting up the experiments; the users need to be properly instructed and the tasks need to be prepared so that the experiment reflects real-world conditions as closely as possible. Furthermore, one needs to take into account the high variability of user behaviour, which is present especially in crowdsourced environments.

*Automated evaluation procedures* A procedure which satisfies the aforementioned requirements has not yet been developed. Most evaluation procedures either require a degree of human involvement in order to be somewhat correlated to human judgement, or they require significant engineering effort. The evaluation methods, which we cover in this survey, can be categorized as follows: model the human judges, model the user behaviour, or use fine-grained methods, which evaluates a specific aspect of the dialogue system (e.g. its ability to adhere to a topic). Methods that model human judges rely on human judgements to be collected beforehand so as to fit a model which predicts the human rating. User behaviour models involve a significant engineering step in order to build a model which emulates the human behaviour. The finer-grained methods also need a certain degree of engineering, which depends on the feature being evaluated. The common trait of the evaluation methods covered in this survey is that they are coupled to the characteristics of the dialogue system that are being considered. That is, a task-oriented dialogue system is evaluated differently to a conversational dialogue system.

### 2.3 Modular structure of this article

Different evaluation procedures have been proposed based on the characteristics of the dialogue system class. For instance, the evaluation of task-oriented systems exploits the highly structured dialogues. The goal can be precisely defined and measured to compute the task-success rate. On the other hand, conversational agents generate dialogues that are more unstructured, which can be evaluated on the basis of appropriateness of the responses; this has been shown to be difficult to automate. We introduce each type of dialogue system to highlight the respective characteristics and methods used to implement the dialogue system. With this knowledge, we introduce the most important concepts and methods developed to evaluate the respective class of dialogue system. In the following survey, we discuss each of the three classes of dialogue systems separately. Thus, Sect. 3: *Task Oriented Dialogue Systems*, Sect. 4: *Conversational Agents*, and Sect. 5: *Interactive Question Answering* can be read independently from each other.

## 3 Task oriented dialogue system

### 3.1 Characteristics

As the name suggests, a task-oriented dialogue system is developed to perform a clearly defined task. These dialogue systems are usually characterized by a clearly defined and measurable goal, a structured dialogue behaviour, a closed domain to work on and a focus on efficiency. Usually, the task involves finding information within a

| Event | Time | Date | Party |
|---|---|---|---|
| trau. appt. | 9am | March 15th | brother |
| lunch | 1pm | March 16th | Cathy |
| trau. appt. | 5pm | April 1st | Mike |
| trau. appt. | 10am | April 4th | Lisa |
| ... | ... | ... | ... |

**DRIVER:** I need to find out the time and parties attending my traumatologist appointment.

**CAR:** I have 3 appointments scheduled, with your brother, Mike and Lisa. Which are you referring to?

**DRIVER:** I want to know about the one that my brother is joining me at.

**CAR:** That appointment is at 9am

**Fig. 1** Example dialogue where the driver can query the agenda via a voice command (Eric et al. 2017). The dialogue system guides the driver through the various options

**Table 2** Domain ontology of the E2E dataset (Novikova et al. 2017)

| Slot | Type | Example values |
|---|---|---|
| ame | Verbatim string | Alimentum,... |
| eatType | Dictionary | restaurant, pub, coffee shop |
| familyFriendly | Boolean | yes, no |
| food | Dictionary | Italian, French, English,... |
| near | Verbatim string | Burger King |
| area | Dictionary | riverside, city center |
| customerRating | Dictionary | 1 of 5, 3 of 5, 5 of 5, low, average, high |
| priceRange | Dictionary | < £ 20, £ 20–25, > £ 30 cheap, moderate, high |

There are eight different slots (or attributes), each has a type and a set of values it can take

database and returning it to the user, performing an action, or retrieving information from its users. For instance, a restaurant information dialogue system helps the user to find a restaurant which satisfies the user's constraints. Furthermore, task-oriented dialogue systems also serve as interfaces to program APIs, which is often used in the Smart Home setting (Möller et al. 2004). For example, an in-car entertainment dialogue system can be ordered to start playing music via voice commands or querying the agenda (see Fig. 1 for an example).

The commonality is that the dialogue system infers the task constraints through the dialogue and retrieves the information requested by the user. For a ticket reservation system, the dialogue system needs to know the origin station, the destination, and the departure date and time. In most cases, the dialogue system is designed for a specific domain, such as restaurant information. The nature of these dialogue systems makes the dialogues both very structured and tailored. The ideal dialogue satisfies the user goal with as few interactions as possible. The dialogues are characterized by mixed initiatives, the user states its goal but the dialogue system proactively asks questions to retrieve the required constraints.

**Table 3** Dialogue acts proposed by Young et al. (2010)

| Dialogue act | Description |
|---|---|
| hello($a = x, b = y, \ldots$) | Open a dialogue and give info $a = x, b = y, \ldots$ |
| inform($a = x, b = y, \ldots$) | Give information $a = x, b = y, \ldots$ |
| request($a, b = x, \ldots$) | Request value for $a$ given $b = x, \ldots$ |
| reqalts($a = x, \ldots$) | Request alternative with $a = x, \ldots$ |
| confirm($a = x, b = y, \ldots$) | Explicitly confirm $a = x, b = y, \ldots$ |
| confreq($a = x, \ldots, d$) | Implicitly confirm $a = x, \ldots$ and request value of $d$ |
| select($a = x, a = y$) | Select either $a = x$ or $a = y$ |
| affirm($a = x, b = y$) | Affirm and give further info $a = x, b = y, \ldots$ |
| negate($a = x$) | Negate and give corrected value $a = x$ |
| deny($a = x$) | Deny that $a = x$ |
| bye() | Close a dialogue |

**Table 4** Sample dialogue and corresponding dialogue acts

| Utterance | Dialogue Act |
|---|---|
| U: Hi, I am looking for somewhere to eat | hello(task = find,type=restaurant) |
| S: You are looking for a restaurant. What type of food? | confreq(type = restaurant,food) |
| U: I'd like an Italian somewhere near the museum. | inform(food = Italian,near=museum) |
| S: Roma is a nice Italian restaurant near the museum. | inform(name = "Roma", type = restaurant, food = Italian, near = museum) |
| U: Is it reasonably priced? | confirm(pricerange = moderate) |
| S: Yes, Roma is in the moderate price range. | affirm(name = "Roma", pricerange = moderate) |
| U: What is the phone number? | request(phone) |
| S: The number of Roma is 385456. | inform(name = "Roma", phone = "385456") |
| U: Ok, thank you goodbye. | bye() |

### 3.2 Dialogue structure

The dialogue structure for task-oriented systems is defined by two aspects: the content of the conversation and the strategy used within the conversation.

*Content* The content of the conversation is derived from the domain ontology. The domain ontology is usually defined as a list of slot-value pairs. For instance, Table 2 shows the domain ontology for the restaurant domain (Novikova et al. 2017). Each slot has a type and a list of values, which the slot can be filled with.

*Strategy* While the domain ontology defines the content of the dialogue, the strategy to fill the required slots during the conversation is modelled as a sequence of actions (Austin 1962). These actions are so-called *dialogue acts*. A dialogue act is defined by its type (e.g. inform, query, confirm, and housekeeping) and by the list of arguments it can take. Each utterance corresponds to an action performed by an interlocutor.

Table 3 shows the dialogue acts proposed by Young et al. (2010).

For instance, the *inform* act is used to inform the user about its arguments, i.e. inform(food = "French", area = "riverside") informs the user that there is a French

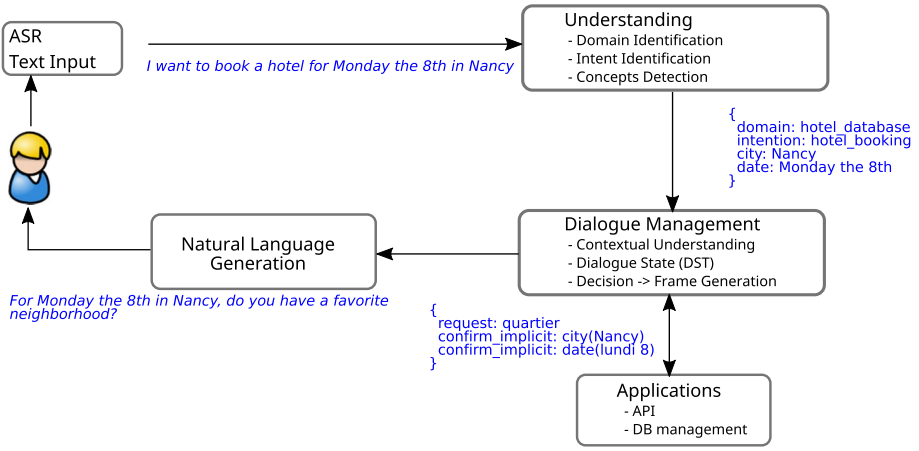


**Fig. 2** General overview of a task-oriented dialogue system

restaurant at the riverside area. On the other hand, the *request* act is used to request a value for a given list of slot-value pairs.

Table 4 shows an example dialogue with the corresponding dialogue acts. Each user utterance is translated into a dialogue act, and each dialogue act of the dialogue system is translated into an utterance in natural language. For instance, the utterance "Hi, I am looking for somewhere to eat" corresponds to the act of "hello". The parameters describe the task that the user intends to solve, i.e. find a restaurant. For a formal description of dialogue acts, refer to Traum (1999); Young (2007).

### 3.3 Technologies

We have just seen that content and strategy are the two main aspects driving the structure of a dialogue, but their influence reaches down to the different functionalities making a classic dialogue system architecture. It is composed of several parts which are built around the idea of modelling the dialogue as a sequence of actions.

The central component is the so-called *dialogue manager*. It defines the dialogue policy, which consists in deciding which action to take at each dialogue turn. The input to the dialogue manager is the current state of the conversation. The output of the dialogue manager is a dialogue act, which represents the system's action. Other components convert the user's input into a dialogue act and the dialogue manager's output into a natural language utterance.

Usually, the user's input is processed by a natural language understanding (NLU) unit, which extracts the slots and their values from the utterance and identifies corresponding the dialogue act. This information is passed to the dialogue state tracker (DST), which infers the current state of the dialogue. Finally the output of the dialogue manager is passed to a natural language generation (NLG) component.

Traditionally, these components were assembled into a pipelined architecture, but recent approaches based on trainable end-to-end neural networks offer a promising alternative. In the following, we briefly introduce the modules of the pipelined architecture and the deep neural network based approach.

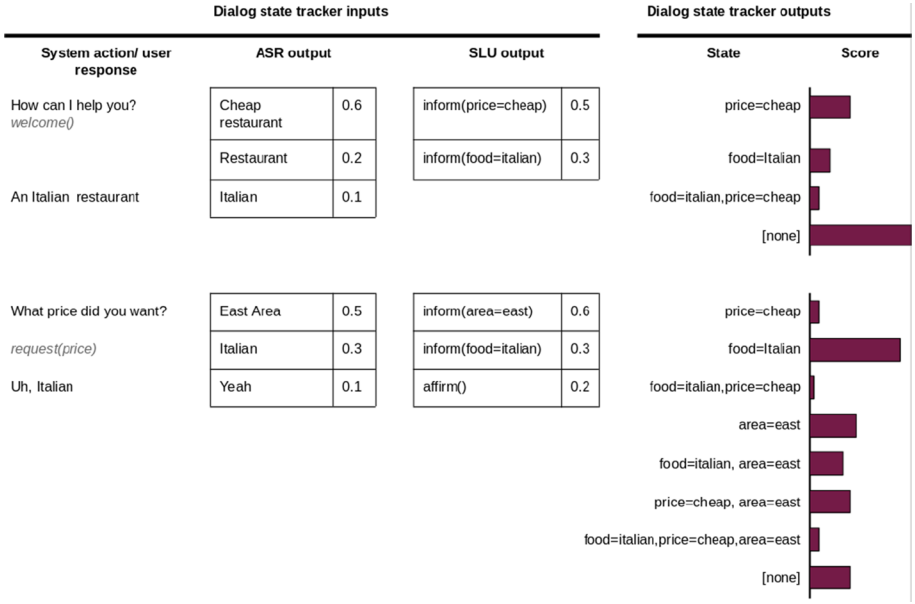


**Fig. 3** Overview of a DST module. The input to the DST module is the combined output of the ASR and the NLU model

### 3.3.1 Pipelined systems

Usually, these four components are put into a pipelined architecture, where the output of one component is fed as the input into the next component (see Fig. 2). The input of the dialogue system is either a chat-interface or an automatic speech recognition (ASR) system. The input to the NLU unit is the utterance of the user in text format or, in the case of automatic speech recognition (ASR) a list of the N-best last user utterance transcriptions.

*Natural language understanding* The goal of the natural language understanding (NLU) unit is to detect the slot-value pairs expressed in the current user utterance. Since the early 2000s, the natural language understanding task is often seen as a set of subtasks (Tur and Mori 2011) as follows: (i) identification of domain (if multiple domains), (ii) identification of intents (that is, the question type, the dialogue act, etc.) and (iii) identification of the slots or concept detection.

In an utterance such as, "I want to book a hotel room for Monday, 8th", the domain is *hotel*, the intent *hotel booking* and the slot-value pair is *date(Monday, 8th)*. The first two tasks are formalized as a classification task and any classification methods may be used. For concept detection one makes use of sequence labelling methods such as *Conditional Random Field* (CRF) (Hahn et al. 2010) or recurrent neural network, typically bi-LSTM with CRF layer (Yao et al. 2014; Mesnil et al. 2015). Recent methods propose to jointly learn the tasks of intent identification and concept detection (Guo et al. 2014; Zhang and Wang 2016). Usually, NLU is performed on classifying the intents that lie within the domain for which the dialogue system is developed for. Larson et al. (2019) introduce an out-of-scope intent classification task, where the NLU system is trained to detect if a user intent does not lie within the scope of the dialogue systems' capabilities.

*Dialogue state tracking* The Dialogue State Tracker (DST) infers the current *belief state* of the conversation, given the dialogue history up to the current point $t$ (Williams et al. 2016). The current belief state encodes the user's goal (e.g. which price range the user prefers) and the relevant dialogue history, i.e. it is an internal representation of the state of the conversation. It is important to take the previous belief states into account in order to handle misunderstandings. For instance, in Fig. 3, the confidence that the user wants an Italian restaurant is low. In the successive turn, the ASR system still gives low confidence to the Italian restaurant. However, since the state tracker takes into account that the Italian restaurant could have been mentioned in the previous turn, it assigns a higher overall probability to it.

The main challenge for the DST module is to handle the uncertainty, which stems from the errors made by the ASR module and the NLU unit. Typically, the output of the DST unit is represented as a probability distribution over multiple possible dialogue states $b(s)$, which provides a representation of the uncertainty. Generative methods have been widely used to manage this task, for example, dynamic Bayesian network (DBN) along with a beam search (Young et al. 2007). Those methods present some limits which are widely discussed in Metallinou et al. (2013), the most important being that all the correlations in the input features have to be modeled (even the unseen cases).

Discriminative models were then proposed to overcome these limits. Metallinou et al. (2013) proposed to use a linear classifier with the dialogue history present in the input features. Whereas Henderson et al. (2013b) proposed to map directly the ASR hypotheses onto a dialogue state by means of recurrent neural networks. This way, both NLU and DST were integrated into a single function. Nowadays, neural approaches are becoming more and more popular (Mrkšić et al. 2017).

*Strategy* The strategy is learned by the dialogue manager. The input is the current belief state $b(s)$ computed by the DST module. The DM generates the next action of the system, which is represented as a dialogue act. In other words, based on the current turn values and on the value history the system performs an action (e.g. retrieve data from a database, ask for a missing information, etc.). Deciding which action to take is part of the dialogue control.

In earlier systems, the dialogue control was based on a finite-state automaton in which the nodes represent the questions of the system and the transitions the possible user's answers. This method, while being rigid, is efficient when the domain and the task are simple. It has been widely used to design dialogue systems and many toolkits are available such as the one from the Center for Spoken Language Understanding (Cole 1999) or VoiceXML.[6] The main issue is the rigid dialogue structure as well as the tendency to be error-prone. In fact, such a system does not model discourse phenomena like ellipsis (a part of the sentence structure that can be inferred from the context is omitted) or anaphoric references (which can be resolved only in a given context).

To overcome these inefficiencies, a dialogue manager is designed to keep track of the interaction history and controls the dialogue strategy. This is called frame-based dialogue control and management. Frame-based techniques rely on schemas specifying what the system has to solve instead of representing what the system has to do and when. This allows for dialogue to be more flexible and the possibility to handle errors (McTear et al. 2005; van Schooten et al. 2007).

[6] See https://www.w3.org/TR/voicexml20/.

Initially, dialogue managers were implemented using rule-based approaches. When data had become available in sufficient amount, data-driven methods were proposed for learning dialogue strategies from data. The dialogue is represented as a Markov decision problem (MDPs), following the intuition that a dialogue can be represented as a sequence of actions (Levin et al. 1998; Singh et al. 2000). These actions are referred to as *speech acts* or *dialogue acts* (Austin 1962; Searle 1969, 1975). However, MDPs cannot handle uncertainty coming from speech recognition errors (Young et al. 2013).

Thus, partially observable MDPs (POMDP) were adopted, as they introduce the belief state, which models the uncertainty of the current state (Paek 2006; Lemon and Pietquin 2012; Young et al. 2013). Although this alleviated the problem of hand-crafting the dialogue policy, the domain ontology still needs to be manually created. Furthermore, these dialogue systems are trained on a static and well-defined domain, once trained the policy works only on this domain. Finally, the dialogue systems need large amounts of data to be trained efficiently, mostly using user simulation for training (Schatzmann et al. 2006). Beyond user simulations, Gašić et al. (2011) showed that online policy learning based on crowdsourcing is a valid alternative.

To mitigate the issues arising from the lack of data, Gašić et al. (2011) applied Gaussian processes for POMDP-based optimization (Engel et al. 2005), which exploits the correlation between different belief states and speeds up the learning process. The authors showed that a reasonable policy can be learned with online user feedback after a few hundred dialogues. Gasic et al. (2013, 2014) showed that it is possible to adapt the policy if the domain is extended dynamically. Note also the work of Wang et al. (2015) which aims at enabling domain-transfer by introducing a domain-independent ontology parametrisation framework.

*Natural language generation* The natural language generation (NLG) module translates the dialogue act represented in a semantic frame into an utterance in natural language (Rambow et al. 2001). The task of NLG is usually divided into separate subtasks such as content selection, sentence planning, and surface realization (Stent et al. 2004). Traditionally, the task has been solved by relying on rule-based methods and canned texts. Statistical methods were also proposed and used, such as phrase-based NLG with statistical language models (Mairesse et al. 2010) or CRF based on semantic trees (Dethlefs et al. 2013). Recently, deep learning techniques have become more prominent for NLG. With these techniques, there now exists a large variety of different network architectures, each addressing a different aspect of NLG; Wen et al. (2015) propose an extension to the vanilla LSTM (Hochreiter and Schmidhuber 1997) to control the semantic properties of an utterance, whereas Hu et al. (2017) use variational autoencoder (VAE) and generative adversarial networks to control the generation of texts by manipulating the latent space; Mei et al. (2016) employ an encoder-decoder architecture extended by a coarse-to-fine aligner to solve the problem of content selection; Wen et al. (2016) apply data counter-fitting to generate out-of-domain training data for pretraining a model where there is little in-domain data available; Semeniuta et al. (2017) and Bowman et al. (2016) use a VAE trained in an unsupervised fashion on large amounts of data to sample texts from the latent space; and Dušek and Jurcicek (2016) use a sequence-to-sequence model with attention to generate natural language strings as well as deep syntax dependency trees from dialogue acts.

### 3.3.2 End-to-end trainable systems

Traditionally, task-oriented dialogue systems were designed along the pipelined architecture, where each module has to be designed, trained, and evaluated separately. There

are several drawbacks to this approach. As the architecture is modular, each component needs to be designed separately, which involves lots of hand-crafting, the costly generation of annotated data for each module, and training each component (Wen et al. 2017). Furthermore, the pipelined architecture leads to the propagation and amplification of errors through the pipeline as each module depends on the output of the previous module (Li et al. 2017b; Liu et al. 2018).

Related to the architecture there is a credit assignment problem, as the dialogue system is evaluated as a whole, it is hard to determine what module is responsible for which reward. Furthermore, this architecture leads to interdependence among the modules, i.e. when one module is changed, all the subsequent modules need to be adapted as well (Zhao and Eskenazi 2016).

Finally, the slot-filling architecture, which is often used, makes these systems inherently hard to scale to new domains since there is a need to hand-craft the representation of the state and action space (Bordes et al. 2017).

To overcome these limitations, current research focuses on end-to-end trainable architectures where the dialogue system is trained as a single module. Wen et al. (2017) model the dialogue as a sequence to sequence mapping, where the traditional pipeline elements are modelled as interacting neural networks. The policy network takes as input the results form the intent network, the belief tracker network, the database operator and selects the next action, based on the selected action, the generation network produces the output utterance.

Bordes et al. (2017) propose a set of synthetic tasks to evaluate the feasibility of end-to-end models in the task-oriented setting, for which they use a memory network to model the conversation. These approaches learn the dialogue policy in a supervised fashion from the data. In contrast the work by Li et al. (2017b); Zhao and Eskenazi (2016) train the system using reinforcement-learning. Note that all these approaches rely on huge amounts of training data.

### 3.4 Evaluation

The evaluation of task-oriented dialogue systems is built around the structured nature of the interaction. Two main aspects are evaluated, which have been shown to define the quality of the dialogue: task-success and dialogue efficiency. Two main metrics of evaluation methods have been proposed:

- *User satisfaction modelling* Here, the assumption is that the usability of the system can be approximated by the satisfaction of its users, which can be measured by questionnaires. These approaches aim to model the human judgements, i.e. creating models which give the same ratings as the human judges. First, a human evaluation is performed where subjects interact with the dialogue system. Afterwards, the dialogue system is rated via questionnaires. Finally, the ratings are used as target labels to fit a model based on objectively measurable features (e.g. task success rate, word error rate of the ASR system).
- *User simulation* Here, the idea is to simulate the behaviour of the users. There are two applications of user simulation: firstly, to evaluate a functioning system with the goal of finding weaknesses and secondly, the user simulation is used as an environment to train a reinforcement-learning based system. The evaluation in the latter is based on the reward achieved by the dialogue manager under the user simulation.

$$C_0 = \begin{bmatrix} type = bar \\ drinks = beer \\ area = central \end{bmatrix}$$

$$R_0 = \begin{bmatrix} name = \\ addr = \\ phone = \end{bmatrix}$$

| attribute | actual value |
|---|---|
| depart-city | Torino |
| arrival-city | Milano |
| depart-range | evening |
| depart-time | 8pm |

**Fig. 4** Examples of goals from Schatzmann et al. (2007) and Walker et al. (1997). Where $C_0$ denotes the information constraints, i.e. which information is to be retrieved (a bar that serves beer in the city center). $R_0$ denotes the set of requests, i.e. the information the user wants (name, address, and phone number)

**Table 5** Confusion matrix from Walker et al. (1997)

| DATA | DEPART-CITY | | | | ARRIVAL-CITY | | | | DEPART-RANGE | | DEPART-TIME | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 | v9 | v10 | v11 | v12 | v13 | v14 |
| KEY | | | | | | | | | | | | | | |
| v1 | **22** | | 1 | | 3 | | | | | | | | | |
| v2 | | **29** | | | | | | | | | | | | |
| v3 | 4 | | **16** | 4 | | | 1 | | | | | | | |
| v4 | 1 | 1 | 5 | **11** | | | 1 | | | | | | | |
| v5 | | | | | **20** | | | | | | | | | |
| v6 | | | | | | **22** | | | | | | | | |
| v7 | | | | | 1 | 1 | **20** | 5 | | | | | | |
| v8 | | | | | 1 | 2 | 8 | **15** | | | | | | |
| v9 | | | | | | | | | **45** | 10 | | | | |
| v10 | | | | | | | | | 5 | **40** | | | | |
| v11 | | | | | | | | | | | **20** | | 2 | |
| v12 | | | | | | | | | | | 1 | **19** | 2 | 4 |
| v13 | | | | | | | | | | | 2 | | **18** | |
| v14 | | | | | | | | | | | 2 | 6 | 3 | **21** |
| sum | 30 | 30 | 25 | 15 | 25 | 25 | 30 | 20 | 50 | 50 | 25 | 25 | 25 | 25 |

For each key (e.g. depart-city) a confusion matrix is created, which denotes the expected values (row) and the values produced by the dialogue system (columns). The maximum value of each column is represented in bold. For instance, if it was expected that the dialogue system returns the train schedule from Torino to Milano but it confused the depart-city with Verona, then this is counted as an error

Both these approaches rely on measuring task-success rate and dialogue efficiency. Before we introduce the methods themselves, we will go over the ways to measure performance along these two dimensions.

*Task-success rate* The goal or the task of the dialogue can be split into two parts (Schatzmann et al. 2007) (see Fig. 4) as follows:

- Set of Constraints, which define the target information to be retrieved. For instance, the specifications of the venue (e.g. a bar in the central area, which serves beer) or the travel route (e.g. ticket from Torino to Milano at 8pm).

- Set of Requests, which define what information the user wants. For instance the name, address and the phone number of the venue.

The task-success rate measures how well the dialogue system fulfills the information requirements dictated by the user's goals. For instance, this includes whether the correct type of venue has been found by the dialogue system and whether the dialogue system returned all the requested information. One possibility to measure this is via a confusion matrix (see Table 5), which represents the errors made over several dialogues. Based on this representation, the Kappa coefficient (Carletta 1996) can be applied to measure the success (see Powers (2012) for Kappa shortcomings).

*Dialogue efficiency* Dialogue efficiency or dialogue costs are measures which are related to the length of the dialogue (Walker et al. 1997) . For instance, the number of turns or the elapsed time are such measures. More intricate measures could include the number of inappropriate repair utterances or the number of turns required for a sub-dialogue to fill a single slot.

In the following, we introduce the most important research for both of the aforementioned evaluation procedures. Finally, we briefly cover the evaluation methods employed on the subsystems of the pipleline. However, the main focus of this review is the evaluation of the dialogue system's behaviour.

### 3.4.1 User satisfaction modelling

User satisfaction modelling is based on the idea that the usability of a system can be approximated by the satisfaction of its users. The research in this area is concerned with three goals: measure the impact of the properties of the dialogue system on the user satisfaction (explainability requirement), automate the evaluation process based on these properties (automation requirement), and use the models to evaluate different dialogue strategies (differentiability requirement). Usually, a predictive model is fit, which takes the properties as input and uses the human judgements as target variable. Thus, modelling the user satisfaction as either a regression or a classification task. There are different approaches to measure the user satisfaction, which are based on two questions: who evaluates the dialogue and at which granularity is the dialogue evaluated? The first question allows for two groups; either the dialogue is evaluated by the users themselves or by objective judges. The second question allows for different points on a spectrum. On one end, the evaluation takes place on the dialogue level, on the other end the evaluation takes place at the exchange level. The question of who evaluates the dialogue is often especially at the centre of discussion. Here, we will shortly summarize the main points.

*User or expert ratings* There are three main criticisms regarding the judgments made by users:

- *Reliability* Evanini et al. (2008) state as a main argument that users tend to interpret the questions on the questionnaires differently, thus making the evaluation unreliable. Gašić et al. (2011) noted that also in the lab setting, where users are given a predefined goal, users tend to forget the task requirements, thus, incorrectly assessing the task success. Furthermore, in the in-field setting, where the feedback is given optionally, the judgements are likely to be skewed towards the positive interactions.

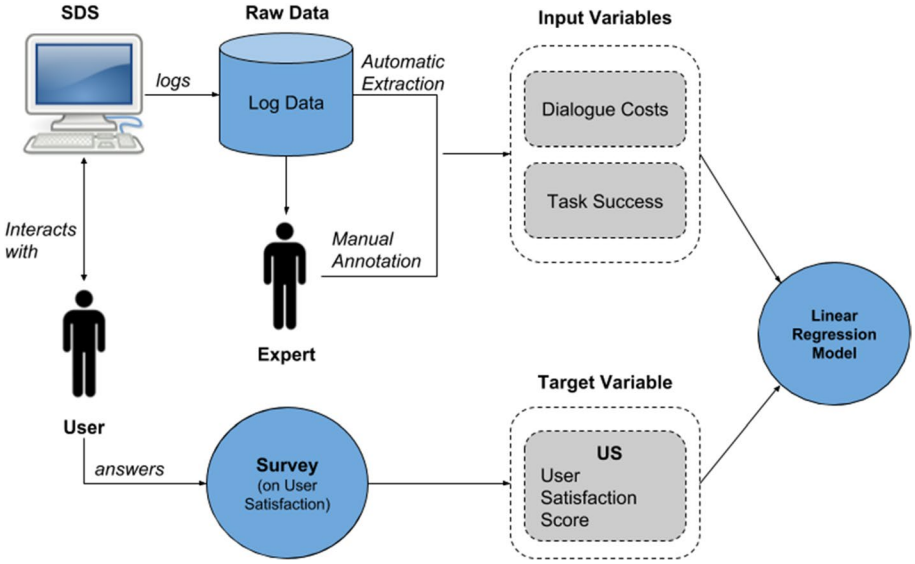


**Fig. 5** PARADISE overview (Schmitt and Ultes 2015)

- *Cognitive demand* Schmitt and Ultes (2015) note that rating the dialogue puts more cognitive demand on users. This is especially true if the evaluation has to be done at the exchange level. This would falsify the judgments about the interaction.
- Impracticability: Ultes et al. (2013) note the impracticability of having a user rate the live dialogue, as he would have to press a button on the phone, or have a special installation to give feedback.

Ultes et al. (2013) analyzed the relation between the user ratings and ratings given by objective judges (called *experts*). Especially, they investigated if the ratings from the experts could be used to predict the ratings of the users. Their results showed that the user ratings and the expert ratings are highly correlated with a Spearman's $\rho$ score of $\rho = 0.66 (p < 0.01)$. Thus, expert ratings can be used as replacement for user judgments. Furthermore, they trained classifiers using the expert rating as targets and evaluated on the user ratings as targets. The best performing classifier achieved an unweighed average recall (UAR) of 0.34 compared to the best classifier trained on user satisfaction, which achieved $UAR = 0.5$. These results indicate that it is not possible to precisely predict the user satisfaction. However the correlation scores show that the predicted scores of both models correlate equally to the user satisfaction $p = 0.6$. Although the models cannot be used to exactly predict the user satisfaction, the authors showed that the expert ratings are strongly related to user ratings.

In the following, we present different approaches to user satisfaction modelling. We cover the most important research for each of the various categories.

**PARADISE Framework** PARADISE (PARAdigm for DIalog System Evaluation) (Walker et al. 1997) is the most known evaluation framework proposed for task-oriented systems. It is a general framework, which can be applied to any task-oriented system, since it is domain-independent. It belongs to the evaluation methods which are based on user ratings on the dialogue level, although it allows for evaluations of sub-dialogues.

**Table 6** Predictive power of PARADISE

| Training set | $R^2$ training (SE) | Test set | $R^2$ test (SE) |
|---|---|---|---|
| ALL 90% | 0.47 (0.004) | ALL 10% | 0.50 (0.035) |
| ELVIS 90% | 0.42 | TOOT | 0.55 |
| ELVIS 90% | 0.42 | ANNIE | 0.36 |
| NOVICES | 0.47 | ANNIE EXPERTS | 0.04 |

Where ALL denotes that the collection of all the annotated data from the three different systems. The distinction between NOVICES and EXPERTS denotes the level to which the test subjects were instructed to use the dialogue system

Originally, the motivation was to produce an evaluation procedure, which can distinguish between different dialogue strategies. At that time, the most widely used automatic approach was based on the comparison of utterances with a reference answer (Hirschman et al. 1990). Methods based on comparisons to reference answers suffer from various drawbacks: they cannot discriminate between different strategies, they are not capable of attributing the performance on system specific properties, and the approach is not generalizable to other tasks.

The main idea of PARADISE is to combine different measures of performance into a single metric, and in turn assess the contribution of each of these measures to the final user satisfaction. PARADISE originally uses two objective measures for performance: task-success and measures that define the dialogue cost (as explained above).

An overview of the PARADISE framework is depicted in Fig. 5. The user interacts with the dialogue system and completes a questionnaire after the dialogue ends. From the questionnaire, a user satisfaction score is computed, which is used as the target variable. The input variables to the linear regression models are extracted from the logged conversation data. The extraction can be done automatically (e.g. for task-success as discussed above) or manually by an expert (e.g. for inappropriate repair utterances). Finally, a linear regression model is fitted to predict the user satisfaction for a given set of input variables.

Thus, PARADISE models the (subjective) performance of the system with a linear combination of objective measures (task-success and dialogue costs). Applying multiple linear regressions showed that only the task-success measure and the number of repetitions are significant. In a follow-up study (Walker et al. 2000), the authors further investigated PARADISE's ability to generalize to other systems and user populations and its predictive power. For this, they applied PARADISE on three different dialogue systems: ELVIS (a dialogue system for accessing emails), ANNIE (a dialogue system for voice dialing and messaging), and TOOT (a dialogue system for accessing train schedules). In a large-scale user study, they collected 544 dialogues over 42 h of speech. For these experiments, the authors worked with an extended number of quality measures: e.g. number of barge-ins (i.e. sudden interruption by the user), number of cancel operations, number of help requests. A survey at the end of the dialogue was used to measure the user satisfaction. The survey asked about various aspects: e.g. speech recognition performance, ease of the task, if the user would use the system again. Based on the survey, the user satisfaction score is computed and used as the target variable to train the PARADISE framework as described above. Table 6 shows the generalization scores of PARADISE for different scenarios.

According to these scores, we obtain the following observations:

- A linear regression model is fitted on 90% of the data and evaluated on the remaining 10%. The results show that the model is able to explain $R^2 = 50\%$ of the variance, which is considered to be a good predictor by the authors.
- Training the regression model on the data for one system and evaluating the model on the data for another dialogue system (e.g. train on the ELVIS data and evaluate on the TOOT data) show high variability as well. The evaluation on the TOOT system data yields much higher scores than evaluating on the ANNIE data. These results show that the model is able to generalize to data of other dialogue systems to a certain degree.
- The evaluation of the generalizability of the model across different populations of users yields a negative result. When trained on dialogue data from conversation by novice users (NOVICES), the linear model is not capable of predicting the scores by experienced users (ANNIE EXPERTS) of the dialogue system.

The PARADISE framework is not only able to find the factors, which have the most impact on the rating, it is also capable of predicting the ratings. However, the experiments also revealed that the framework is not capable of distinguishing between different user groups. This result was confirmed by Engelbrecht et al. (2008), which tested the predictive power of PARADISE for individual users.

*User satisfaction at the exchange level* In contrast to rating the dialogue as a whole, in some cases it is important to know the rating at each point in time. This is especially useful for online dialogue breakdown detection. There are two approaches to modelling the user satisfaction at the exchange level: annotate dialogues at the exchange level either by users (Engelbrecht et al. 2009a) or by experts (Higashinaka et al. 2010; Schmitt and Ultes 2015). Different models can be fitted with the sequential data: Hidden Markov Models (HMM), Conditional Random Fields or Recurrent Neural Networks are the most obvious choice, but also SVM based approaches are possible.

Engelbrecht et al. (2009a) model user satisfaction as a continuous process evolving over time, where the current judgment depends on the current dialogue events and the previous judgments. Users interacted with the dialogue system and judged the dialogue after each turn on a 5-point scale using a number pad. An HMM was trained based on these target values and annotated dialogue features. Some input features were manually annotated, which is not a reasonable setting for online breakdown detection.

Higashinaka et al. (2010) modelled the evaluation similarly as in Engelbrecht et al. (2009a). In their study, they evaluated different models (HMM and CRF), different measures to evaluate the trained model, and addressed the question of subjectivity of the annotators. The input features to the model were the dialogue acts and the target variables were the annotations by experts, which listened to the dialogue. The low inter-rater agreement and the fact of only using dialogue acts as inputs made the model perform only marginally better than the random baseline.

A different approach was taken by Hara (2010), who relied on dialogue-level ratings, but trained the model on n-grams of dialogue-acts. More precisely, they used as input features $n$ consecutive dialogue acts and used the dialogue-level rating as target variable (on a 5-point scale and an extra class to denote unsuccessful task). The model achieved an accuracy of only 34.4% using a 3-gram model. Further testing yielded that the model is able to predict the task-success with an accuracy of 94.7%.

These approaches suffer from the following problems: they either rely on manual feature extraction, which is not useful for online breakdown detection or they used only dialogue acts as input features, which does not cover the whole dialogue complexity. Furthermore, the approaches had issues with data annotation, either having low inter-rater agreement

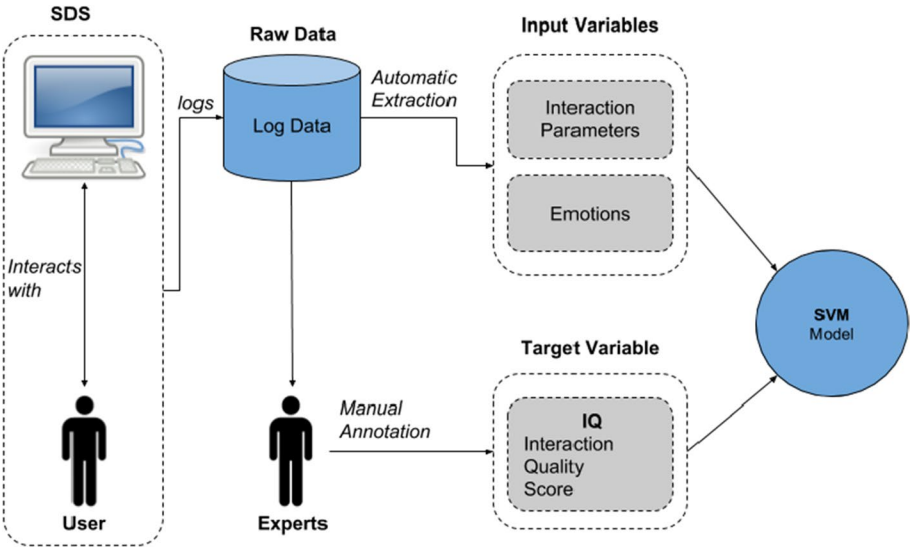


**Fig. 6** Overview of the interaction quality procedure (Schmitt and Ultes 2015)

or using dialogue-level annotation. Schmitt and Ultes (2015) addressed these issues by proposing Interaction Quality (see next paragraph) as approximation to user ratings at the exchange level.

*Interaction quality* Interaction Quality is a metric proposed by Schmitt and Ultes (2015) with the goal to allow the automatic detection of problematic dialogue situations. The approach is based on letting experts rate the quality of the dialogue at each point in time—the median rating of several expert ratings at the exchange level is called Interaction Quality. The experiments in this study were conducted using the *Let's Go* bus information system Black and Eskenazi (2009).

Figure 6 shows the overview of the Interaction Quality procedure. The user interacts with the dialogue system and the conversation's relevant data is logged. From the logs, the input variables are automatically extracted. The target variables are manually annotated by experts, from which the target variable is derived. Based on the input and target variables, a support vector machine (SVM) is fitted.

Interaction Quality is meant to approximate user satisfaction. In this study, the authors showed that Interaction Quality is an objective and valid approximation to user satisfaction, which is easier to obtain. This is especially important for in-field evaluations of dialogue systems, which are practically infeasible to be rated by users at the exchange level. Thus, it is important that in-field dialogues can be rated by experts at the exchange level. The challenge is to make sure that the ratings are objective, i.e. to eliminate the subjectivity of the experts as much as possible.

Since there is no possibility to gather user satisfaction scores at the exchange level from in-field conditions, the authors relied on user satisfaction scores from lab experiments and Interaction Quality scores over dialogues from both in-field and lab conditions. For the lab experiments, users interacted with the *Let's Go* bus information system (Black and Eskenazi 2009) and used a special device to rate the dialogue after each turn. These scores are referred to as user satisfaction. The dialogues were then rated by experts on the exchange level. These ratings are referred to as Interaction Quality. The authors found a

**Table 7** Model performance (in terms of $\rho$) on the test set. Schmitt and Ultes (2015)

| Feature set | $IQ_{field}$ | $IQ_{lab}$ | $US_{lab}$ |
|---|---|---|---|
| ASR | 0.753 | 0.811 | 0.625 |
| AUTO | 0.776 | 0.856 | 0.668 |
| AUTO + EMO | 0.785 | 0.856 | 0.669 |
| AUTO + EMO + USER | – | 0.888 | 0.741 |

ASR denotes the features by the automatic speech recognition system. AUTO denotes automatically extracted features from the dialogue system pipeline (e.g. dialogue acts). EMO denotes features that capture the users emotions (e.g. anger). USER denotes user specific features (e.g. age, gender)

**Table 8** Model performance (in terms of $\rho$, $\kappa$ and UAR) on the test set Schmitt and Ultes (2015)

| Feature set | Test | Train | $\rho$ |
|---|---|---|---|
| Auto | $US_{lab}$ | $IQ_{lab}$ | 0.667 |
| Auto | $IQ_{lab}$ | $IQ_{field}$ | 0.647 |
| Auto | $IQ_{field}$ | $IQ_{lab}$ | 0.696 |

strong correlation (Spearman's $\rho = 0.66$) between Interaction Quality and user satisfaction in the lab environment, which means that Interaction Quality is a valid substitute for user satisfaction. In order to assess if Interaction Quality is a valid measure for rating in-field conversations, experts rated 200 dialogues from the *Let's Go Field Corpus* (Schmitt et al. 2012) and measured the agreement among the experts. The experts achieved a strong correlation (Spearman's $\rho = 0.72$).

Based on these Interaction Quality scores a predictive model is trained to automatically judge the dialogue at any point in time. In order to automatically predict Interaction Quality, the input variable need to be automatically extractable from the dialogue system. From each subsystem of a task-oriented dialogue system (Fig. 2), various values are extracted (AUTO features). Additionally, the authors experimented with hand-annotated features such as emotions (EMO) and user specific features (USER), such as age or gender, as well as semi-automatically annotated data such as the dialogue acts (similar to Higashinaka et al. 2010). Based on these input variables, the authors trained various SVMs, one for each target variable, namely Interaction Quality for both in-field and the lab data as well as the user satisfaction label for the lab data. Table 7 shows the scores achieved for the various target variables and input feature groups.

The in-field Interaction Quality model ($IQ_{field}$) achieves a correlation of $\rho = 0.776$ to the human judges, based on the automatically extracted features, with the ASR features alone the correlation score lies at $\rho = 0.753$. The addition of the emotional and user -specific features do not increase the scores significantly. A similar behaviour is measured for the lab Interaction Quality model ($IQ_{lab}$), which achieves high scores with ASR features alone ($\rho = 0.856$) and profits only marginally from the inclusion of the emotional features. However, the model improves when including user specific features ($\rho = 0.894$). The lab based user satisfaction model ($US_{lab}$) achieves lower scores with $\rho = 0.668$ for the automatic features.

Table 8 shows the cross model evaluation. The $IQ_{field}$ model can be used to predict $IQ_{lab}$ labels and vice versa ($\rho \sim 0.66$). Furthermore, the $IQ_{lab}$ model is able to predict the $US_{lab}$ variable. These results show that Interaction Quality is a good substitute to user satisfaction

and that the models based on Interaction Quality yield high predictive performance when trained on the automatically extracted features. This allows to evaluate an ongoing dialogue in real-time at the exchange level and ensures high correlation to the actual user satisfaction.

### 3.4.2 User simulation

User Simulators (US) are tools that are designed to simulate the user's behaviour. There are two main applications for US: (1) for training the dialogue manager in an offline environment, and (2) to evaluate the dialogue policy.

*Training environment* User Simulations are used as a learning environment to train reinforcement -learning based dialogue managers. They mitigate the problem of recruiting humans to interact with the systems, which is both time- and cost-intensive. There is a vast amount of literature on designing User Simulations as training environment, for a comprehensive survey refer to Schatzmann et al. (2006). There are several considerations to be made when building a User Simulation.

- *Interaction level* Does the interaction take place at the semantic level (i.e. on the level of dialogue acts) or at the surface level (i.e. using natural language understanding and generation)?
- *User goal* Does the simulation update the goal during the conversation or not? The dialogues in the second Dialogue State Tracking Challenge (DSTC2) data contain a large amount of examples where the user changes their goal during the interaction (Henderson et al. 2014). Thus, it is more realistic to model these changes as well.
- *Error model* Whether and how to realistically model the errors made by the components of the dialogue system.
- *Evaluation of the user simulation* For a discussion on this topic refer to Pietquin and Hastie (2013). There are two main evaluation strategies: direct and indirect evaluation. The direct evaluation of the simulation is based on metrics (e.g. precision and recall on dialogue acts, perplexity). The indirect evaluation measures the utility of the user simulation (e.g. by evaluating the trained dialogue manager).

The most popular approach to user simulation is based on the agenda-based user simulation (ABUS) (Schatzmann et al. 2007). The simulations takes place at the semantic level, the user goal stays fixed throughout the interaction, and the user behaviour is represented as a priority ordered stack of necessary user actions. The ABUS was evaluated using indirect methods, by performing a human study on a dialogue system trained with the ABUS. The results show that the DS achieved an average task success rate of 90.6% based on 160 dialogues. The ABUS system works by randomly generating a hidden user goal (i.e. the goal is unknown to the dialogue system), which consists of constraints and request slots. From this goal, the ABUS system generates a stack of dialogue acts in order to reach the goal, which is the agenda. During the interaction with the dialogue system, the ABUS adapts the stack after each turn, e.g. if the dialogue system misunderstood something, the ABUS system pushes a negation act onto the stack.

Similar to other aspects of dialogue systems, more recent work is based on neural network based approaches. The Neural User Simulator (NUS) by (Kreyssig et al. 2018) proposes an end-to-end trainable architecture based on neural networks. The system performs the interaction on the surface instead of the semantic level, during the training it considers

variable user goals, and the evaluation is performed indirectly. The indirect evaluation is performed from two different perspectives. First, the dialogue system, which is trained with the NUS is compared to a dialogue system trained with ABUS in the context of a human evaluation. Here, the authors report the average reward and the success rate. In both cases the NUS-trained system performs significantly better. The second evaluation is performed in a cross-model evaluation (Schatztnann et al. 2005), i.e. the NUS-trained dialogue system is evaluated using the ABUS system and vice-versa. Here, the NUS system performed significantly better as well. This indicates that the NUS system is diverse and realistic.

*Model based evaluation* The idea of model based evaluation is to model the user behaviour but to put more emphasis on modelling a large variety of behavioural aspects. Here, the focus does not lie in the shaping of rewards for reinforcement learning, rather, the focus lies on understanding the effects of different types of behaviour on the quality of the interaction. Furthermore, the goal is to gain insights on the effects of adapting a dialogue strategy, i.e. evaluate the changes made to the dialogue system. Engelbrecht et al. (2009b) introduced the MeMo workbench, which allows the modelling of user simulations. The main focus is to model different types of users and typical errors the users make. Möller et al. (2006) introduced various types of conceptual errors, which users tend to make. There errors arise from the discrepancy between how the user expects the system to behave and the actual system behaviour. For instance:

- State errors arise when the user input cannot be interpreted in the current state, but might be interpretable in a different state.
- Capability errors arise when the system cannot execute the user's commands due to missing capability.
- Modelling errors arise due to discrepancies in how the user and the system model the world. For instance, when presented with a list of options and the system allows to address the elements in the list by their positions, but the user addresses them by their name.

On the other hand, the workbench allows the definition of various user groups based on different characteristics of a user. The characteristics used in Engelbrecht et al. (2009b) include: affinity to technology, anxiety, problem solving strategy, domain expertise, age and deficits (e.g. hearing impairment). Behavioural rules are associated to each of the characteristics. For instance, a user with high domain expertise might use a more specific vocabulary. The rules are manually curated and are engineered to influence the probabilities of user actions. During the interaction, the user model selects a task to solve similar to the aforementioned approaches for reinforcement-learning environments. In order to evaluate the user simulation, the authors compared the results of an experiment conducted with real users to the experiments conducted with the MeMo workbench. This evaluation procedure is aimed at finding whether the simulation yields the same insights as a user study. For this, they invited users from two user groups, namely older and younger users. The participants interacted with two versions of a smart-home device control system: the versions differed in the way they provide help to the users. The comparison between the user simulation and the user study results was done at various levels:

- High-level features, such as concept error rates or average number of semantic concepts per user turn (#) AVP. Here, the results show that the simulation was not always able to recreate the absolute values, it was able to replicate the relative results. This is helpful, as it would lead to the same conclusions for the same questions.

- User judgment prediction which is based on a predictive model trained using the PARADISE framework. Here, the authors compared the real user judgments to the predicted judgments (where the linear model predicted the judgments of the simulated dialogue). Again, the results show that the user model would yield the same conclusions as the user study, namely that young users rated the system higher than the older users and that old users judged the dynamic help system worse than the other.
- Precision and Recall of predicted actions. Here, the simulation is used to predict the next user action for a given context from a dialogue corpus. The predicted user action is compared to the real user action and based on this precision and recall is computed. The results show that precision and recall are relatively low.

The model-based user simulations are designed with the idea of allowing the evaluation of a dialogue system early in the development stage. Furthermore, they emphasize the need of interpretability, i.e. being able to understand how a certain change in the dialogue system influences the quality of the dialogue. This lies in contrast to the user simulations for reinforcement learning, which are aimed at training a dialogue system and use the reward as a measure of quality. However, the reward is often only based on the task success and the number of turns.

### 3.4.3 Subsystems evaluation

This section briefly outlines the different evaluation metrics employed on every subsystem, composing a pipelined Dialogue System, namely Natural Language Understanding, Dialogue State Tracker and Natural Language Generation systems.

*Natural language understanding (NLU)* Since NLU is often cast as a classification task, NLU systems are often evaluated in the literature with regard to classification-based metrics. There are three widely used metrics (Tur and De Mori 2011): Sentence Level Semantic Accuracy (SLSA), Slot Error Rate (SER) (also called Concept Error Rate (CER)), and F-measures. The SLSA measures the rate of sentences where the intents are correctly classified. The SER metric measures the rate of inserted, deleted or substituted concepts with respect to the annotated concept as a reference. Finally, the F-measures compute the precision and recall of the correctly detected slots. In early systems, the distance between hypothesized sentences and reference ones is calculated with a Levenshtein distance (Levenshtein 1966) or using the Word Error Rate (Chotimongkol and Rudnicky 2001), which fail to capture the semantic similarities of utterances.

*Dialogue state trackers (DST)* DST usually report a probability distribution over the possible next states. In order to measure the performance of such systems, accuracy and L2 metrics are widely used (Metallinou et al. 2013; Henderson et al. 2014; Mrkšić et al. 2017). Accuracy measures whether the state hypothesis with the higher probability is the correct one. Having a high accuracy is crucial because DST systems must commit to a single interpretation of user's needs. L2 metric captures how well calibrated the output probabilities are, which is important when multiple dialogue states are considered in action selection.

*Natural language generation (NLG)* NLG systems translate the dialogue act into natural language, the dialogue act is composed of slot-value pairs, which the NLG system renders. The evaluation focuses on two aspects: the correctness of the content and the quality of the surface realization. For the correctness, the F1 score is used (Mei et al. 2016), as well as the slot error rate (Wen et al. 2015) (i.e. the ratio of the slots which have been correctly rendered). For the quality of the surface realization, the word overlap metrics are used (e.g.

BLEU (Papineni et al. 2002), or ROUGE (Lin 2004)). However, since the automated metrics do not necessarily capture all aspects of the output's quality, usually a human evaluation is performed, which usually asks about the naturalness and quality of the generated utterance (Dušek et al. 2020).

## 4 Conversational dialogue systems

### 4.1 Characteristics

Conversational dialoge systems (also referred to as chatbots and social bots) are usually developed for unstructured, open-domain conversations with its users. They are often not developed with a specific goal in mind, other than to maintain an engaging conversation with the user (Zhou et al. 2018). These systems are usually built with the intention to mimic human behaviour, which is traditionally assessed by the Turing Test (more on this later). However, Conversational dialogue systems might also be developed for practical applications. "Virtual Humans", for instance, are a class of conversational agents developed for training or entertainment purposes. They mimic certain human behaviours for specific situations. For instance, a Virtual Patient mimics the behaviour of a patient, which is then used to train medical students (Kenny et al. 2009; Mazza et al. 2018). Early versions of conversational agents stem from the psychology community with ELIZA (Weizenbaum 1966) and PARRY (Colby 1981). ELIZA was developed to mimic a Rogerian psychologist, whereas PARRY was developed to mimic a paranoid mind.

*Modelling approaches* Generally, there are two main approaches for modelling a Conversational dialogue system: *rule-based systems* and *corpus-based systems*.

Early systems, such as *ELIZA* (Weizenbaum 1966) and *PARRY* (Colby 1981) are based on a set of rules which determine their behaviour. ELIZA works on pattern recognition and transformation rules, which take the user's input and apply transformations to it in order to generate responses.

Recently, conversational dialogue systems have gained a renewed attention in the research community, as shown by the recent effort to generate and collect data for the (RE-)WOCHAT workshops.[7] This renewed attention is motivated by the opportunity of exploiting large amounts of dialogue data (see Serban et al. (2018) for an extensive study as well as Sect. 6) to automatically author a dialogue strategy that can be used in conversational systems such as chatbots (Banchs and Li 2012; Charras et al. 2016). Most recent approaches train conversational agents in and end-to-end fashion using deep neural networks, which mostly rely on the sequence-to-sequence architecture (Sutskever et al. 2014).

In the following, we focus on the corpus-based approaches used to model conversational agents. First, we describe the general concepts, and then the technologies used to implement conversational agents. Finally, we cover the various evaluation methods which have been developed in the research community.

### 4.2 Modelling conversational dialogue systems

Generally, there are two different strategies to exploit large amounts of data:

[7] See http://workshop.colips.org/re-wochat/ and http://workshop.colips.org/wochat/.

- *Utterance selection* Here, the dialogue is modelled as an information retrieval task. A set of candidate utterances is ranked by relevance. The dialogue structure is thus defined by the utterances in a dialogue database (Lee et al. 2009). The idea is to retrieve the most relevant answer to a given utterance, thus learning to map multiple semantically equivalent user-utterances to an appropriate answer.
- *Generative models* Here, the dialogue systems are based on deep neural networks, which are trained to generate the most likely response to a given conversation history. Usually, the dialogue structure is learned from a large corpus of dialogues. Thus, the corpus defines the dialogue behaviour of the conversational agent.

Utterance selection methods can be interpreted as an approximation to generative methods. This approach is often used for modelling the dialogue system of Virtual Humans. Usually, the dialogue database is manually curated and the dialogue system is trained to map different utterances of the same meaning to the same response utterance. Another application of utterance selection is applied to integrate different systems (Serban et al. 2017b; Zhou et al. 2018). Here, the utterance selection system selects from a candidate list, which is comprised of outputs of different subsystems. Thus, given a set of dialogue systems, the utterance selection module is trained to select for the given context, the most suitable output from the various dialogue systems. This approach is especially interesting for dialogue systems, which work on a large number of domains and incorporate a large amount of skills (e.g. set alarm clock, report the news, return the current weather forecast). Here, we present the technologies for corpus-based approaches, namely the neural generative models and the utterance selection models.

#### 4.2.1 Neural generative models

The architectures are inspired by the machine translation literature (Ritter et al. 2011), especially neural machine translation. Neural machine translation models are based on the Sequence to Sequence (seq2seq) architecture (Sutskever et al. 2014), which is composed of an encoder and a decoder. They are usually based on a Recurrent Neural Network (RNN). The encoder maps the input into a latent representation on which the decoder is conditioned. Usually, the latent representation of the encoder is used as the initial state of the recurrent cell in the decoder. The earliest approaches were proposed by Shang et al. (2015); Vinyals and Le (2015), which trained a seq2seq model on a large amount of dialogue data (in the order of $10^6$ exchanges). There are two fundamental weaknesses with the neural conversational agents. Firstly, they do not take into account the context of the conversation. Since the encoder only reads the current user input, all previous states are ignored. This leads to dialogues, where the dialogue system does not refer to previous information, which might lead to nonsensical dialogues. Secondly, the models tend to generate generic answers that follow the most common pattern in the corpus. This renders the dialogue monotonous and in the worst case leads to repeating the same answer, regardless of the current input. We briefly discuss these two aspects in the following section.

*Context* The context of the conversation is usually defined as the previous turns in the conversations. It is important to take these into account as they contain information relevant to the current conversation. Sordoni et al. (2015) proposed to model the context by adding the dialogue history as a bag-of-words representation. The decoder is then conditioned on the encoded user utterance and the context representation. An alternative approach was

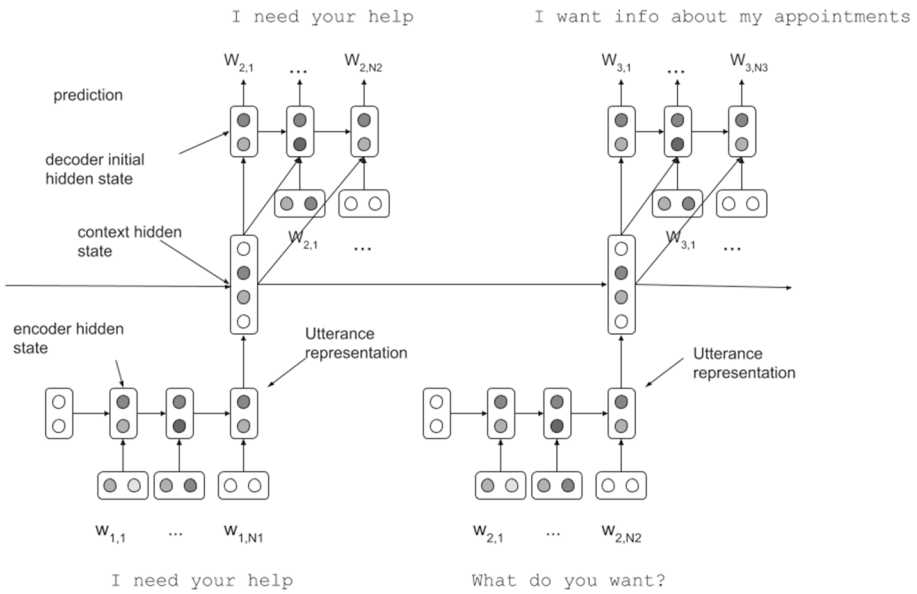


**Fig. 7** Overview of the HRED architecture. There are two levels of encoding: (i) the utterance encoder, which encodes a single utterance and (ii) the context encoder, which encodes the sequence of utterance encodings. The decoder is conditioned on the context encoding

proposed by Serban et al. (2016), who proposed the hierarchical-encoder decoder architecture (HRED), shown in Fig. 7, which works in three steps:

1. A turn-encoder (usually a recurrent neural network) encodes each of the previous utterances in the dialogue history, including the last user utterance. Thus, for each of the preceding turns a latent representation is created.
2. A context-encoder (a recurrent neural network) takes the latent turn representations as input and generates a context representation.
3. The decoder is conditioned on the latent context representation and generates the final output.

The HRED architecture is used as basis for more complex neural architectures for dialogue system, such as the multi-resolution recurrent neural network (MrRNN) (Serban et al. 2017a), which extends the HRED architecture by adding encoders that capture different levels of granularity (e.g. entity level, word level, or action level). Furthermore, the HRED encoder is used to generate the representation for the context in the *utterance selection models* (see Sect. 4.2.2).

*Variability* There are two main approaches on dealing with the issue of repetitive and universal responses:

- Adapt the loss functions. The main idea is to adapt the loss function in order to penalize generic responses and promote more diverse responses. Li et al. (2016a) propose two loss functions based on maximum mutual information: one is based on an anti-language model, which penalizes high-frequency words; the other is based on the probability of

the source given the target. Li et al. (2016b) propose to train the neural conversational agent using the reinforcement-learning framework. This allows to learn a policy that can plan in advance and generate more meaningful responses. The major focus is the reward function, which encapsulates various aspects: ease of answering (reduce the likelihood of producing a dull response), information flow (penalize answers that are semantically similar to a previous answer given), and semantic coherence (based on the mutual information).

- Condition the decoder. The seq2seq models perform a *shallow* generation process. This means that each sampled word is only conditioned on the previously sampled words. There are two methods for conditioning the generation process: condition on stochastic latent variables or on topics. Serban et al. (2017c) enhance the HRED model with stochastic latent variables at the utterance level and on the word level. At the decoding stage, first the latent variable is sampled from a multivariate normal distribution and then the output sequence is generated. Xing et al. (2017) add a topic-attention mechanism in their generation architecture, which takes as inputs *topic words* which are extracted using the Twitter LDA model (Zhao et al. 2011). The work by Ghazvininejad et al. (2018) extends the seq2seq model with a *Facts Encoder*. The "facts" are represented as a large collection of raw texts (Wikipedia, Amazon reviews, etc.), which are indexed by named entities.

### 4.2.2 Utterance selection methods

Utterance selection methods generally try to devise a similarity measure that measures the similarity between the dialogue history and the candidate utterances. There are roughly three different types of such measures:

- Surface form similarity. This measures the similarity at the token level. This includes measures such as: Levenshtein distance, METEOR (Lavie and Denkowski 2009), or TF-IDF retrieval models (Charras et al. 2016; Dubuisson Duplessis et al. 2016). For instance, Dubuisson Duplessis et al. (2017) propose an approach that exploits recurrent surface text patterns to represent dialogue utterances.
- Multi-class classification task. These methods model the selection task as a multi-class classification problem, where each candidate response is a single class. For instance, Gandhe and Traum (2013) model each utterance as a separate class, and the training data consists of utterance-context pairs on which features are extracted. Then a perceptron model is trained to select the most appropriate response utterance. This approach is suitable for applications with a small amount ($\sim$ 100) of candidate answers.
- Neural network based approaches. Neural network architectures were introduced to leverage large amounts of training data. Usually, they are based on a siamese architecture, where both the current utterance and a candidate response are encoded. Based on this representation a binary classifier is trained to distinguish between relevant responses and irrelevant. One well-known example is the dual encoder architecture proposed by Lowe et al. (2017b). Dual Encoders transform the user input and a candidate response into a distributed representation. Based on the two representations a logistic regression layer is trained to classify the pair of utterance and candidate response as either relevant or not. The softmax score of the relevant class is used to sort the candidate responses. The authors experimented with different neural network architectures for modelling

the encoder, such as recurrent neural networks or long short-term memory networks (LSTM) (Hochreiter and Schmidhuber 1997).

### 4.3 Evaluation methods

Automatically evaluating conversational dialogue systems is an open problem. The difficulty in automating this step can be attributed to the characteristics of the conversational dialogue system. Without a clearly defined goal or task to solve, and a lack of structure in the dialogues, it is not clear which attributes of the conversation are relevant to measure the system's quality. Two common approaches to assess the quality of a conversational dialogue system are to measure the appropriateness of its responses, or to measure the human likeness thereof. Both these approaches are very coarse-grained and might not reveal the complete picture. Nevertheless, most approaches in evaluation follow these principles. Depending on the characteristics of a specific dialogue system, more fine-grained approaches to evaluation can be applied, which measure the capability of the specific characteristic. For instance, a system built to increase the variability of its answers might be evaluated based on lexical complexity measures (such as token-type ratio or lexical density. For a more in-depth discussion please refer to Lu (2012). In the following, we introduce the automated approaches for evaluating conversational dialogue systems. In the first part, we discuss the general metrics that can be applied to both the generative models as well as the selection-based models. We then survey the approaches specifically designed for the utterance selection approaches, as they can exploit various metrics from information retrieval.

#### 4.3.1 General metrics for conversational dialogue systems

There are generally two levels in order to evaluate a conversational dialogue system: coarse-grained and fine-grained evaluations. The coarse-grained evaluations focus on the adequacy of the responses generated or selected by the dialogue system. On the other hand, fine-grained evaluations focus on specific aspects of its behaviour. Coarse-grained evaluations are based on two concepts: adequacy (or appropriateness) of a response, and the human likeness thereof. Fine-grained evaluations focus on specific behaviours that a dialogue system should manifest. Here, we focus on the methods devised for coherence and the ability of maintaining the topic of a conversation. In the following, we give an overview of the methods that have been designed to automatically evaluate the above dimensions.

*Appropriateness* This is a coarse-grained concept to evaluate a dialogue, as it encapsulates many finer-grained concepts, e.g. coherence, relevance, or correctness, among others. There are two main approaches in the literature: word-overlap based metrics and methods based on predictive models inspired by the PARADISE framework (see Sect. 3.4.1).

- *Word-overlap metrics* These metrics were originally proposed by the machine translation and the summarization community. They were initially a popular choice of metrics for evaluating dialogue systems seeing as they are easily applicable. Popular metrics such as BLEU score (Papineni et al. 2002) and ROUGE (Lin 2004) were used as approximation for the appropriateness of an utterance. However, Liu et al. (2016) showed that neither of the word-overlap based scores have any correlation to human judgments.

Based on the criticism of the word-overlap metrics, several new metrics have been proposed. Galley et al. (2015) propose to include human judgments into the BLEU score, which they call ΔBLEU. The human judges rated the reference responses of the test set according to the relevance to the context. The ratings are used to weight the BLEU score to reward high-rated responses and penalize low-rated responses. The correlation to human judgments was measured by means of Spearman's $\rho$. ΔBLEU has a correlation of $\rho = 0.484$, which is significantly higher than the correlation of the BLEU score, which lies at $\rho = 0.318$. Although this increases the correlation of the metric to the human judgments, this procedure involves human judgments to label the reference sentences.

- *Trained metrics* Lowe et al. (2017a) present an automatic dialogue evaluation model (ADEM), a recurrent neural network trained to predict appropriateness ratings by human judges. The human ratings were collected via Amazon Mechanical Turk, where the judges were presented with a dialogue context and a candidate response, which they rated on appropriateness on a scale from 1 to 5. Based on the ratings, a recurrent neural network was trained to score the model response, given the context and the reference response. The Pearson's correlation between ADEM and the human judgments is computed on two levels: the utterance level and at the system level, where the system level rating is computed as the average score at the utterance-level achieved by the system.

  The Pearson's correlation for ADEM lies at 0.41 on the utterance level and at 0.954 on the system level. For comparison, the correlation to human judgments for the ROUGE score only lies at 0.062 on the utterance level and at 0.268 at the system level.

  While ADEM relies on human labelled data, Tao et al. (2018) present a method, which has no need of human judges. Their model is based on two observations. Firstly, a response that is close to the ground truth is likely to be good. Secondly, a response that is related to the last utterance or the context of the conversation is good. They propose two submodels to capture these insights. The first model computes a representation of both the ground truth and the generated response based on min- and max-pooling of word embeddings. Then the cosine similarity is computed to measure the relatedness of the ground truth and the generated response. The second model rates the relatedness between the conversational context and the generated response. In order to train this model, the authors create a training set of positive examples (i.e. pairs of contexts and responses that are relevant) and negative examples (i.e. pairs of irrelevant contexts and responses). The positive examples are taken from the dialogues in the training material, whereas the negative examples are constructed by randomly sampling utterances from the corpus for a given context. Then a siamese neural network is trained on this training data to predict if a pair of context and response are relevant. The scores of both submodules are then normalized and averaged. The Pearson's correlation for their model lies at 0.4594, which is comparable to ADEM.

  Although trained metrics have a significantly higher correlation to human judgements, they are show not to be robust (Sai et al. 2019). In fact, with simple manipulations of the response under consideration can lead to significant changes in the score of ADEM. For instance, in 48.66% of cases the predicted score increased when the generated response was reversed. In 86.93% of cases the predicted score increased when the generated response was replaced with a dull dummy response. Thus, creating reliable trained metrics is still an open problem.

*Human likeness* The classic approach to measure the quality of a conversational agent is the Turing Test devised by Turing (1950). The idea is to measure if the conversational

dialogue system is capable of fooling a human into thinking that it is a human as well. Thus, according to this test, the main measure is the ability to imitate human behaviour.

Inspired by this idea, the use of *adversarial learning* (Goodfellow et al. 2014) can be applied to evaluate a dialogue system. The framework of a generative adversarial model is composed of two parts: the generator, which generates data, and the discriminator, which tries to distinguish whether the data is real or artificially generated. The two components are trained in an adversarial manner: the generator tries to fool the discriminator, and the discriminator learns at the same time to identify if the data is real or artificial. Adversarial Evaluation of dialogue systems was first studied by Kannan and Vinyals (2016), where the authors trained a generative adversarial network (GAN) on dialogue data, and used the performance of the discriminator as indicator for the quality of the dialogue. The discriminator achieved an accuracy of 62.5% which indicates a weak generator. However, the authors did not evaluate whether the discriminator score is a viable metric for evaluating a dialogue system.

A study on the viability of adversarial evaluation was conducted by Bruni and Fernandez (2017). For this, they compared the performance of discriminators to the performance of humans on the task of discriminating between real and artificially generated dialogue excerpts. Three different domains were used, namely: MovieTriples (46k dialogue passages) (Serban et al. 2016), SubTle (3.2M dialogue passages) (Banchs 2012) and Switchboard (77k dialogue passages) (Godfrey et al. 1992). The GAN was trained on the concatenation of the three datasets. The evaluation was conducted on 900 dialogue passages, 300 per dataset, which were rated by humans as real or artificially generated. The results show that the annotator agreement among humans was low, with a Fleiss (Fleiss 1971) $\pi = 0.3$, which shows that the task is difficult. The agreement between the discriminator and the humans is on par with the agreement among the humans, except for the Switchboard corpus, where $\pi = 0.07$. Human annotators achieve an accuracy score with respect to the ground-truth of 64–67.7% depending on the domain. The discriminator achieves lower accuracy scores on the Switchboard dataset but higher scores than humans on the other two datasets.

In order to evaluate the ability of the discriminators on different models, a seq2seq model was trained on the OpenSubtitles dataset (Tiedemann 2009) (80M dialogue passages). The discriminator and the human performance on the dialogues generated by the seq2seq model was evaluated. The results show that the discriminator performs better than the humans, which the authors attribute to the fact that the discriminators may pick up on patterns that are not apparent to humans. The agreement between humans and the discriminator is very low.

*Fine-grained metrics* The above methods for evaluating conversational dialogue systems work on a coarse-grained level. The dialogue is evaluated on the basis of producing adequate responses or its ability to emulate human behaviour. These concepts encompass more finer-grained concepts. In this section, we look at topic-based evaluation.

*Topic-based evaluation* This measures the ability of a conversational agent to talk about different topics in a cohesive manner. Guo et al. (2018) propose two dimensions of topic-based evaluation: topic breadth (can the system talk about a large variety of topics?) and topic depth (can the system sustain a long and cohesive conversation about one topic?). For topic classification, a Deep Averaging Network (DAN) was trained on a large amount of question data. DANs do topic classification and the detection of topic-specific keywords. The conversational data used to evaluate the topic-based metrics

stems from the Alexa-Prize challenge,[8] which consists of millions of dialogues and hundreds of thousands of live user ratings (on a scale from 1 to 5). Using the DAN, the authors classified the dialogue utterances according to the topics.

Conversational *topic depth* is measured by the average length of a sub-conversation on a specific topic, i.e. multiple consecutive turns where the utterances are classified as being the same topic. The conversational breadth is measured on a coarse- and fine-grained level. Coarse-grained topic breadth is measured as the average number of topics a bot converses about during a conversation. On the other hand, *topic breadth* measures looks at the total number of distinct topic keywords across all conversations.

To measure the validity of the proposed metrics, correlations between the metric and the human judgments are computed. The conversational topic depth metric has a correlation of $\rho = 0.707$ with the human judgments. The topic breadth metric has a correlation of $\rho = 0.512$ with the human judgments. The lower correlation of the topic breadth is attributed to the fact that the users may not have noticed a bot repeating itself as they only conversed with a bot a few times.

#### 4.3.2 Utterance selection metrics

The evaluation of dialogue systems based on utterance selection differs from the evaluation of generation-based dialogue systems. Here, the evaluation is based on metrics used in information retrieval, especially Recall@k (R@k). R@k measures the percentage of relevant utterances among the top-k selected utterances. One major drawback of this approach is that potentially correct utterances among the candidates could be regarded as incorrect.

*Next Utterance selection* Lowe et al. (2016) evaluate the impact of this limitation and evaluate whether the Next Utterance Classification (NUC) task is suitable to evaluate dialogue systems. For this, they invited 145 participants from Amazon Mechanical Turk (AMT) and 8 experts from their lab. The task was to select the correct response given a dialogue context (of at most six turns) and five candidate utterances, of which exactly one is correct. Note that the other four utterances could also be relevant, but are regarded as incorrect in this experiment. The study was performed on dialogues of three different domains: the SubTle Corpus (Banchs 2012) consisting of movie dialogues, the Twitter Corpus (Ritter et al. 2010) consisting of user dialogues, and the Ubuntu Dialogue Corpus (Lowe et al. 2015), which consists of conversations about Ubuntu related topics.

The human performance was compared to the performance of an artificial neural network, which is trained to solve the same task. The performance was measured by means of R@1 score. The results show that for all domains, the human performance was significantly above random, which indicates that the task is feasible. Furthermore, the results show that the human performance varies depending on the domain and the expertise level. In fact, the lab participants performed significantly better on the Ubuntu domain, which is regarded as harder as it requires expert knowledge. This shows that there is a range of performance that can be achieved. Finally, the results showed that the ANN achieved similar performance to the human non-experts and performed worse than the experts. This shows that this task is not trivial and by far not solved. However, the authors did not take into account the fact that multiple candidates responses could be regarded as correct. This is possible since the selection of the candidate response is performed by sampling at random from the corpus. On the other hand, it is

[8] https://developer.amazon.com/alexaprize.

not clear if their evaluation suffered from this potential limitation, as their results showed the feasibility and relevance of the NUC task.

DeVault et al. (2011) and Gandhe and Traum (2016) tackle the problem of having multiple relevant candidate utterances and propose a metric which takes this into account. Their metrics are both dependent on human judges and measure the appropriateness of an utterance.

*Weak agreement*DeVault et al. (2011) propose the *weak agreement* metric. This metric is based on the observation that human judges only agree in about 50% of the cases on the same utterance for a given context. The authors attribute this to the fact that multiple utterances could be regarded as acceptable choices. Thus, the weak agreement metric regards an utterance as appropriate if at least one annotator chose this utterance to be appropriate.

The authors apply the weak agreement metric on the evaluation of a Virtual Human which simulates a witness in a war-zone and is designed to train military personnel in Tactical Questioning (Gandhe et al. 2009). They gathered 19 dialogues and 296 utterances in a Wizard-of-Oz experiment. To allow for more diversity, they let human experts write paraphrases of the commander role to ensure that the virtual character understands a larger variety of inputs. Furthermore, the experts expanded the set of possible answers by the virtual character by annotating other candidate utterances as appropriate.

The weak agreement metric was able to measure the improvement of the system when the extended dataset was applied: the simple system based on the raw Wizard-of-Oz data achieved a weak agreement of 43%; augmented with the paraphrases, the system achieved a score of 56%; and, finally, adding the manual annotation increases the score to 67%. Thus, the metric is able to measure the improvements made by the variety in the data.

*Voted appropriateness* One major drawback of the weak agreement is that it depends on human annotations and is not applicable to large amounts of data. Gandhe and Traum (2016) improve upon the idea of weak agreement by introducing the *Voted Appropriateness* metric. Voted Appropriateness takes the number of judges into account which selected an utterance for a given context. In contrast to weak agreement, which regarded each adequate utterance equally, Voted Appropriateness weights each utterance.

Similarly to the PARADISE approach, the authors of Voted Appropriateness fit a linear regression model on the pairs of utterances and contexts labelled with the amount of judges that selected the utterance. The fitted model only explains 23.8% of the variance. The authors compared the correlation of the Voted Appropriateness and the weak agreement metric to human judgments. The correlation was computed on the individual utterance level and the system level. For the system level, the authors used data from seven different dialogue systems and averaged the ratings over all dialogues of one system. On the interaction level, the Voted Appropriateness achieved a correlation score of 0.479 ($p < 0.001$, $n = 397$), and the weak agreement achieved 0.485 ($p < 0.001$, $n = 397$). On the system level, Voted Appropriateness achieved 0.893 ($p < 0.01$, $n = 7$) and weak agreement achieved 0.803 ($p < 0.001$, $n = 397$). Thus, on the system level Voted Appropriateness performs closer to human judgments. Both metrics rely heavily on human annotations, which makes the metrics hardly suitable for large-scale data driven approaches.

## 5 Question answering dialogue systems

A different form of task-oriented systems are Question Answering (QA) systems. Here, the task is defined as finding the correct answer to a question. This setting differs from the aforementioned task-oriented systems in the following ways:

- Task-oriented systems are developed for a multitude of tasks (e.g. restaurant reservation, travel information system, virtual assistant, etc.),whereas QA systems are developed to find answers to specific questions.
- Task-oriented systems are usually domain-specific, i.e. the domain is defined in advance through an ontology and remains fixed. In contrast, QA systems usually work on broader domains (e.g. factoid QA can be done over different domains at once), although there are also some QA systems focused only on a specific domain (Sarrouti and Ouatik El Alaoui 2017; Do et al. 2017).
- The dialogue aspect for QA systems is not tailored to sound human-like, rather, the focus is set on the completion of the task. That is, to provide a correct answer to the input question.

### 5.1 Characteristics

Generally, QA systems allow the users to search for information using a natural language interface, and return short answers to the user's question (Voorhees 2006). QA systems can be broadly categorized into three categories (Bernardi and Kirschner 2010): single-turn QA, context QA, and Interactive QA.

*Single-turn QA* Single-turn QA is the most common type of system. Here, the system is developed to return a single answer to the users' question without any further interaction. These systems work very well for factoid questions (Voorhees 2006). However, they have difficulties handling complex questions, which require several inference steps (Iyyer et al. 2017a) or situations where systems need additional information from the user (Li et al. 2017a).

Single-turn QA can be approached from two main perspectives (Rogers et al. 2020a):

- Open QA, where systems collect evidences and answers across several sources such as Web pages and knowledge bases (Fader et al. 2013)
- Reading Comprehension (RC), where the answer is gathered from a single document. This is the most common approach.

RC systems can be oriented to:

- Extractive RC, where systems extract spans of text with the answer. This approach has received a lot of attention fostered by the availability of popular benchmarks such as SQuAD (Rajpurkar et al. 2018), NewsQA (Trischler et al. 2017) or TriviaQA (Joshi et al. 2017). Each of these datasets contains thousands of examples, which permits to train Deep Learning systems and obtain good results.
- Multiple-choice RC, where systems must select an answer from a set of candidates. Multiple-Choice (MC) is a common way to measure reading comprehension in humans. This is why some researches have pointed MC as a better format to test language understanding of automatic systems (Rogers et al. 2020a). There exists several MC collections, mostly in English. In some cases it involves paying crowd-workers to gather documents and/or pose questions regarding those documents. MCTest (Richardson et al. 2013), for example, proposed for the workers to invent short, children friendly, fictional stories and four questions with four answers each, including deliberately wrong answers. As a way to encourage a deeper understanding of texts, the QuAIL dataset includes unanswerable questions (Rogers et al. 2020b). Other datasets were created from real world exams. This is the case of the well known MC dataset RACE (Lai et al. 2017), or the multilingual Entrance Exams (Rodrigo et al. 2018).

- Generative QA, where systems create a text that answers the question. The exact text is not necessarily contained in any document, which makes this a challenging task. This kind of systems has received less attention given that it is difficult to perform an exact evaluation and there are few datasets available (Kočiský et al. 2018).

There is a large amount of research in the area of single-turn QA and there are several surveys, we refer the reader to: Kolomiyets and Moens (2011); Diefenbach et al. (2018); Mishra and Jain (2016). In this survey, we focus on the evaluation of multi-turn QA systems, which is a much less researched area.

*Context QA* Context QA refers to systems which allow for follow-up questions to resolve ambiguities or keeping track of a sequence of inference steps (Peñas et al. 2012). The questions can be highly context-dependent and elliptical, with references to previous questions and answers, which can be seen as a dialog. In fact, it is common to include pronouns instead entities. That is, systems must rely not only on the source document and last question, but also on the context given by previous questions and answers.

Context QA systems are also named multi-turn QA (Choi et al. 2018) or sequential QA (Saha et al. 2018). The most common approach is to develop these systems for extractive RC. In some cases, context QA systems are used for answering complex questions. These systems assume that some complex questions are usually unrealistic but they can be decomposed into simpler inter-related questions (Iyyer et al. 2017b). Then, the system answers the simpler questions and obtain an answer to the initial complex question (Talmor and Berant 2018).

*Interactive QA* Interactive QA (IQA) systems combine context QA systems and task-oriented dialogue systems. The main purpose of the conversation module is to handle under-or-over constrained questions (Qu and Green 2002). E.g. if a question does not yield any results, the system might propose to relax some constraints. In contrast, if a question yields too many results, the interaction can be used to introduce new constraints to filter a list of results (Rieser and Lemon 2009). For a more in-depth discussion on IQA systems, refer to Konstantinova and Orasan (2013).

### 5.2 Technologies

Current QA technologies for single-turn QA are based on pre-trained transformer models such as BERT (Devlin et al. 2019), XLNet (Yang et al. 2019) or ALBERT (Lan et al. 2019). These models have been pre-trained from unlabeled text to do Masked Language Modeling and Next Sentence Prediction. Afterwards, each model can be fine-tuned in specific tasks such as those at Glue (Wang et al. 2018) or QA.

Fine-tuning for QA systems is done by modelling the span detection as prediction of the start and end token in the paragraph. The input to the system is a pair of question and paragraph. Thus, the trained system will output the span with the highest probability of being an answer to the question. These systems achieve the best results for extractive QA, as it can be seen in the corresponding leaderboards of the most popular collections[9]

In the case of multi-turn QA, systems must be aware of the dialogue history. One approach is to reuse single-turn systems, augmenting the input with previous questions and answers (Huang et al. 2019). In some cases, the system may focus on modelling information gain and include pre-trained models such as BERT (Yeh and Chen 2019).

[9] https://rajpurkar.github.io/SQuAD-explorer/.

Given the importance of dealing with answer history, other researchers have proposed to represent answer history using embeddings from pre-trained models (Qu et al. 2019). Then, the system includes also a history attention mechanism to help in the selection of items in the history of the dialogue.

Other models include Adversarial Training and Knowledge Distillation over ROBERTA (Liu et al. 2019) to perform a better fine-tunning of pre-trained models (Ju et al. 2019). While Adversarial Training allows improving the performance of the system against data perturbations, Knowledge Distillation transfers knowledge from one machine to another to improve results of the second machine (Furlanello et al. 2018).

### 5.3 Evaluation of QA dialogue systems

The evaluation of QA systems has two aspects: the correctness of the answer and the flow of the conversation. Currently, most QA systems are evaluated based on the correctness of their answers. Even for multi-turn QA systems, the dialogue flow is often ignored during evaluation (Reddy et al. 2018; Choi et al. 2018).

*Correctness metrics* The evaluation of QA systems depends on the output of the system. For open QA, where the output is a ranking of sentences with potential answers, the evaluation is mostly based on ranking measures such as Mean Average Precision (MAP) or Mean Reciprocal Rank (MRR), but there are also evaluations based on precision, recall and F1 (Yang et al. 2015).

For multiple-choice RC the task is evaluated using accuracy (Clark and Etzioni 2016), that is, the number of times in which the system selected the correct answer.

For extractive QA, which is the most common approach, the output is a span of text. The retrieved span is compared with the ground truth answers and two kinds of evaluations are given (Rajpurkar et al. 2016):

- Exact matching, which measures the percentage of candidate answers that match any one of the ground truth answers exactly.
- Approximate matching based on F1, which measures the macro-average overlap between the bag of words of candidates and ground truth answers.

*Dialogue evaluation* The nature of multi-turn QA systems makes it quite hard to design accurate evaluation frameworks that go beyond the correctness measures, which do not take into consideration the dialogue aspect of the interaction. In fact, a proper evaluation of multi-turn QA systems requires humans to interact with the systems. The first evaluation framework designed specifically for IQA systems is based on a series of questionnaires to capture different aspects of the system (Kelly et al. 2009). The authors argue that metrics based on the relevance of the answers are not sufficient to evaluate an IQA system (e.g. it does not take the user feedback into account). Thus, they evaluate the usage of different questionnaires in order to assess the different systems. The questionnaires they propose are:

- NASA TLX (cognitive workload questionnaire): Used to measure the cognitive workloads as subjects completed different scenarios.
- *Task questionnaire* After each task the questionnaire is filled out, which focuses on the experiences of using a system for a specific task.
- *System questionnaire* Compiled after using a system for multiple tasks. This measures the overall experiences of the subjects.

Their evaluation showed that the Task Questionnaire is the most effective at distinguishing among different systems.

The evaluation of dialogue QA systems requires one to simulate some interactions with users and evaluate them. These interactions can, on the one hand, be created by real users, which is associated with high costs, and makes it hard to reproduce the experiments and reuse the data. For example, Li et al. (2017c) developed DailyDialog, a multi-turn dataset with 13k dialogues created by humans that also include emotion information.

On the other hand, the interactions can be automatically produced. However, it is challenging to create users' responses automatically. One approach for creating simulations is to provide some feedback based on the supplementary questions. For example, if an additional question asks for a location, the simulator can return a location contained in the dialogue's history (or related to it) (Li et al. 2017a). Nevertheless, the simulation can generate several errors. On the other hand, the simulation might only reward the generation of questions similar to a given template (Li et al. 2016b), which constrains the diversity of questions.

There is usually a weak correlation between automatic evaluations and human judgements in multi-turn QA. This is because most of the current QA dialogue systems are trained and tested using data where there is only a single response for each context (Serban et al. 2017c). Moreover, this data contains only a possible path to reach the correct answer, while the same answer could be reached with a different dialogue. In fact, there are many features involved in deciding the next response in a dialogue. This has been defined as the one-to-many problem of dialogues (Zhao et al. 2017).

Automatic evaluations based on multiple-reference responses have been proposed to alleviate the one-to-many problem. Multi-reference based evaluations include several correct responses for a given context. Thus, these evaluations promote diversity better than single-response approaches.

Sordoni et al. (2015) created a synthetic multiple-reference dialogue corpus based on Twitter. Additional responses to the initial response were searched using Information Retrieval and rated by crowd workers. The authors kept only responses with a high rate. Galley et al. (2015) created a dataset from Twitter following the work from Sordoni et al. (2015). However, Galley et al. (2015) included all the synthetic responses (no matter the rate given by crowd workers) and used the data for testing a new metric called Discriminative BLEU.

Sugiyama et al. (2019) performed another evaluation based on multiple-reference responses. They measured the correlation, using a regression-based approach, between systems' responses and a large set of both positive and negative human references. Gupta et al. (2019) extended the test split of DailyDialog (1k dialogues) with multiple references. They compared the results of using single-reference versus multiple-reference data. Both works showed a higher correlation of automatic evaluations with human judgments when using multiple-reference dialogues instead of single-reference data.

## 6 Evaluation datasets and challenges

Datasets play an important role for the evaluation of dialogue systems, together with challenges open to public participation. A large number of datasets have been used and made publicly available for the evaluation of dialogue systems in the last decades, but the coverage across dialogue components and evaluation methods (e.g. Sects. 3 and 4) is uneven.

**Table 9** Datasets for task-oriented dialogue systems

| Name | Topics | # dialogues | Reference |
|---|---|---|---|
| DSTC1 | Bus schedules | 15,000 | (Williams et al. 2013) |
| DSTC2 | Restaurants | 3000 | (Henderson et al. 2014) |
| DSTC3 | Tourist information | 2265 | (Henderson et al. 2013a) |
| DSTC4 & DSTC5 | Tourist information | 35 | (Kim et al. 2016) |
| DSTC6 | Restaurant reservation | – | (Perez et al. 2017) |
| DSTC7 (Flex Data) | Student guiding | 500 | (Gunasekara et al. 2019) |
| DSTC8 (MetaLWOz) | 47 domains | 37,884 | (Lee et al. 2019) |
| DSTC8 (Schema-Guided) | 20 domains | 22,825 | (Rastogi et al. 2019) |
| MultiWOZ | Tourist information | 10,438 | (Budzianowski et al. 2018) |
| Taskmaster-1 | 6 domains | 13,215 | (Byrne et al. 2019) |
| MultiDoGo | 6 domains | 86,698 | (Peskov et al. 2019) |

**Table 10** Datasets for conversational dialogue systems

| Name | Topics | # dialogues | References |
|---|---|---|---|
| Switchboard | Casual topics | 2400 | Godfrey et al. (1992) |
| British national corpus | Casual topics | 854 | Leech (1993) |
| SubTle corpus | Movie subtitles | 3.35M | Ameixa and Coheur (2013) |
| Reddit domestic abuse corpus | Abuse help | 21,133 | Schrading (2015) |
| Twitter corpus | Unrestricted | 1.3M | Ritter et al. (2010) |
| Twitter triple corpus | Unrestricted | 4322 | Sordoni et al. (2015) |
| Ubuntu dialogue corpus | Ubuntu problems | 930K | Lowe et al. (2015) |
| bAbI | Restaurant reservation | 3000 | Bordes et al. (2017) |
| OpenSubtitles | Movie subtitles | 36M | Tiedemann (2012) |
| CornellMovie | Movie dialogues | 220K | Danescu and Lee (2011) |

**Table 11** Datasets for question answering dialogue systems

| Name | Topics | # dialogues | References |
|---|---|---|---|
| Ubuntu dialogue corpus | Ubuntu problems | 930K | Lowe et al. (2015) |
| MSDialog | Microsoft products | 35K | Qu et al. (2018) |
| ibAbI | Restaurant reservation | – | Li et al. (2017a) |
| CoQA | 7 domains | 8399 | Reddy et al. (2018) |
| QuAC | People | 13,594 | Choi et al. (2018) |
| DoQA | Cooking | 1637 | Campos et al. (2019) |

Also note that datasets are not restricted to specific evaluation methods, as they can be used to feed more than one evaluation method or metric interchangeably. In this section, we cover the most relevant datasets and challenges, starting with select datasets. For further references, refer to a broad survey of publicly available datasets that have already been used to build and evaluate dialogue systems carried out by Serban et al. (2018).

The dialogue datasets selected for this section are listed in Tables 9, 10 and 11, where properties such as the topics covered and number of dialogues are indicated.

### 6.1 Datasets for task-oriented dialogue systems

Datasets are usually designed to evaluate specific dialogue components, and very few public datasets are able to evaluate an entire task-oriented dialogue system (e.g. Sect. 3). The evaluation of these kinds of systems is highly system-specific, and it is therefore difficult to reuse the dataset with other systems. Their evaluation also requires considerable human effort, as the involvement of individual users or external evaluators is usually needed. For example, in Gasic et al. (2013), which is a Partially observable Markov decision process -based dialogue system mentioned in Sect. 3.3.1 for the restaurants domain, the evaluation of policies is done by crowd-sourcers via the Amazon Mechanical Turk service. Mechanical Turk users were asked first to find some specific restaurants, and after each dialogue was finished, they had to fill in a feedback form to indicate if the dialogue had been successful or not. Similarly, for the end-to-end dialogue system by Wen et al. (2017) (cf. Sect. 3.3.2), also for the restaurants domain, human evaluation was conducted by users recruited via Amazon Mechanical Turk. Each evaluator had to follow a given task and to rate the system's performance. More specifically, they had to grade the subjective success rate, the perceived comprehension ability and naturalness of the responses.

Most of the task-oriented datasets are designed to evaluate components of dialogue systems. For example, several datasets have been released through different editions of the Dialog State Tracking Challenge,[10] focused on the development and evaluation of the dialogue state tracker component. However, even if these datasets were designed to test state tracking, Bordes et al. (2017) used them to build and evaluate a whole dialogue system, re-adjusting the dataset by ignoring the state annotation and reusing only the transcripts of dialogues. The Schema Guided Dialogue (SGD) dataset released for the 8th edition of DSTC was designed to test not only state tracking, but also intent prediction, slot filling and language generation for large-scale virtual assistants. SGD consists of almost 23K annotated multi-domain (bank, media, calendar, travel, weather, ...), task-oriented dialogues between a human and a virtual assistant.

The MultiWOZ (Multi-Domain Wizard-of-Oz) dataset represented a significant breakthrough in the scarcity of dialogues as it contains around 10K dialogues, which is at least one order of magnitude larger than any structured corpus available before (Budzianowski et al. 2018). It is annotated with dialogue belief states and dialogue actions, so it can be used for the development of individual components of a dialogue system. But its considerable size makes it very appropriate for the training of end-to-end based dialogue systems. The main topic of the dialogues is tourism, containing seven domains, such as attractions, hospitals, police, hotels, restaurants, taxis and trains. Each dialogue can contain more than one of these domains.

Similar in size and content to MultiWOZ is Taskmaster-1 task-based dialogue dataset (Byrne et al. 2019). It includes around 13K dialogues in six domains: ordering pizza, setting auto repair appointments, arranging taxi services, ordering movie tickets, ordering coffee drinks and making restaurant reservations. What makes it different from the previous

[10] https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge/.

one is that more than a half of the dialogues are created following a self-dialogue methodology, in which a crowd-worker writes the full dialogue themselves. The authors claim that these self-dialogues have richer and more diverse language than, for example, MultiWOZ, as it is not restricted to a small knowledge base.

The largest human-generated and multi-domain dialogue dataset that is available to the public is MultiDoGo (Peskov et al. 2019), which comprises over 81K dialogues. These dialogues were created following the Wizard-of-Oz approach between a crowd-worker and a trained annotator. These participants were guided to introduce specific biases like intent or slot change, multi-intent, multiple slot values, slot overfilling and slot deletion in conversations. Additionally, over 54K of the total amount of the dialogues are annotated at the turn level for intent classes and slot labels. Dialogues are from six different domains: airline, fast food, finance, insurance, media and software support.

We will conclude this section by discussing two related tools, rather than a dialogue dataset. The first tool, called PyDial,[11] partially addresses the shortage of evaluation datasets for task-oriented systems. This is because it offers the opportunity for developing a dialogue management environment, based on reinforcement-learning for benchmarking purposes (Ultes et al. 2017). Thus, it makes it possible to evaluate and compare different task-oriented dialogue systems in the same conditions. This toolkit not only provides domain-independent implementations of different modules in a dialogue system, but also simulates users (see Sect. 3.4.2). It uses two metrics for the evaluation: (1) the average success rate and (2) the average reward for each evaluated policy model of reinforcement-learning algorithms. The success rate is defined as the percentage of dialogues that are completed successfully. Thus, it is closely related to the task-completion metric used by the PARADISE framework (see Sect. 3.4.1).

Another dialogue annotation tool is called LIDA (Collins et al. 2019). The authors argue that the quality of a dataset has a significant effect on the quality of a dialogue system, hence, a good dialogue annotation tool is essential to create the best annotated dialogue dataset. LIDA is the first annotation tool that handles the entire dialogue annotation pipeline from turn and dialogue segmentation through to labelling structured conversation data. Moreover, it also includes an interface for inter-annotator disagreements resolution.

### 6.2 Datasets for conversational dialogue systems

Regarding the evaluation of Conversational dialogue systems presented in Sect. 4, datasets derived from conversations on micro-blogging or social media websites (e.g. Twitter or Reddit) are good candidates, as they contain general-purpose or non-task-oriented conversations that are orders of magnitude larger than other dialogue datasets used before. For instance, Switchboard (Godfrey et al. 1992) (telephone conversations on pre-specified topics), British National Corpus (Leech 1993) (British dialogues many contexts, from formal business or government meetings to radio shows and phone-ins) and SubTle Corpus (Ameixa and Coheur 2013) (aligned interaction-response pairs from movie subtitles) are three datasets released earlier that have 2400, 854 and 3.35 M dialogues and 3 M, 10 M and 20 M words,

[11] http://www.camdial.org/pydial/.

respectively. These sizes are relatively small if we compare to the huge Reddit Corpus[12] which contains over 1.7 billion comments,[13] or the Twitter Corpus described below.

Because of the limit on the number of characters permitted in each message on Twitter, the utterances are quite short, very colloquial and chat-like. Moreover, as the conversations happen almost in real-time, the conversations of this micro-blogging website are very similar to spoken dialogues between humans. There are two publicly available large corpora extracted from Twitter. The former one is the Twitter Corpus presented in Ritter et al. (2010), which contains roughly 1.3 million conversations and 125M words drawn from Twitter. The latter is a collection of 4232 three-step (context-message-response) conversational snippets extracted from Twitter logs.[14] This is labeled by crowdsourced annotators, who measure the quality of a response in a given context (Sordoni et al. 2015).

Alternatively, Lowe et al. (2015) hypothesized that chat-room style messaging is more closely correlated to human-to-human dialogues than micro-blogging websites like Twitter, or forum-based sites such as Reddit. Thus, they presented the above-mentioned Ubuntu Dialogue Corpus. This large-scale corpus targets a specific domain. Thus, it could accordingly be used as a task-oriented dataset for the research and evaluation of dialogue state trackers. However, it also has the unstructured nature of interactions from microblog services that makes it appropriate for the evaluation of non-task-oriented dialogue systems.

These two large datasets are adequate for the three subtypes of non-task-oriented dialogue systems: unsupervised, trained and utterance selection metrics. Notice that, additionally, some human judgments could be needed in some cases, such as in Lowe et al. (2017a) for the ADEM system (see Sect. 4.3.1). Here, they use human judgments collected via Amazon Mechanical Turk in addition to the evaluation using the Twitter dataset.

Apart from the afore-mentioned two datasets, the five datasets generated recently for bAbI tasks (Bordes et al. 2017) are appropriate for evaluation using the next utterance classification method (see Sect. 4.3.2). These tasks were designed for testing end-to-end dialogue systems in the restaurant domain, but they check whether the systems can predict the appropriate utterances among a fixed set of candidates, and are not useful for systems that generate the utterance directly. The ibAbI dataset mentioned in the next section has been created based on bAbI to cover several representative multi-turn QA tasks.

Another interesting resource is the ParlAI framework[15] for dialogue research, as it contains many popular datasets available all in one place with the goal of sharing, training and evaluating dialogue models across many tasks (Miller et al. 2017). Some of the dialogue datasets that are included have been already mentioned (bAbI Dialog tasks and the Ubuntu Dialog Corpus) but it also contains conversations mined from OpenSubtitles[16] and Cornell Movie.[17]

[12] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/.

[13] As far as we know, this dataset has not been used in any research work. Researchers have used smaller and more curated versions of the Reddit dataset like Reddit Domestic Abuse Corpus Schrading (2015), which contains 21,133 dialogues.

[14] https://www.microsoft.com/en-us/download/details.aspx?id=52375.

[15] http://parl.ai/.

[16] http://opus.lingfil.uu.se/OpenSubtitles.php.

[17] https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html.

### 6.3 Datasets for question answering dialogue systems

With respect to QA dialogue systems, two datasets have been created based on human interactions from technical chats or forums. The first one is the Ubuntu Dialogue Corpus, containing almost one million multi-turn dialogues extracted from the Ubuntu chat logs, which was used to receive technical support for various Ubuntu-related problems (Lowe et al. 2015). Similarly, MSDialog contains dialogues from a forum dedicated to Microsoft products. MSDialog also contains the user intent of each interaction (Qu et al. 2018).

ibAbI represents another approach for creating multi-turn QA datasets (Li et al. 2017a). ibAbI interactivity adds to the bAbI dataset that was previously presented (see Sect. 6.2) by adding sentences and ambiguous questions with the corresponding disambiguation question, which should be asked by an automatic system. The authors evaluate their system regarding the successful tasks. However, it is unclear how to evaluate a system if it produces a modified version of the disambiguation question.

Recently, several datasets that are very relevant for the context of QA dialogue systems have been released. The CoQA (Conversational Question Answering) dataset contains 8K dialogues and 127K conversation turns (Reddy et al. 2018). The answers from CoQA are free-form text with their corresponding evidence highlighted in the passage. It is a multi-domain dataset, as the passages are selected from several sources, covering seven different domains: children's stories, literature, middle and high school English exams, news, articles from Wikipedia, science and discussions from Reddit. QuAC (Question Answering in Context) consists of 14K information-seeking QA dialogues (100K total QA pairs) over sections from Wikipedia articles about people (Choi et al. 2018). What makes it different from other datasets so far is that some of the questions are unanswerable and that context is needed in order to answer some of the questions. Another similar dataset that has unanswerable questions and its questions are context-dependent is DoQA, a dataset for accessing domain-specific Frequently Asked Question sites via conversational QA (Campos et al. 2019). It contains 1,637 information-seeking dialogues on the cooking domain (7,329 questions in total). An analysis carried out by the authors showed that in this dataset there are less factoid questions than in the others, as DoQA focuses on open-ended questions about specific topics. Amazon Mechanical Turk was used to collect the dialogues for the three datasets.

### 6.4 Evaluation challenges

We conclude this section by summarizing some of the recent evaluation challenges that are popular for benchmarking state-of-the-art dialogue systems. They have an important role in the evaluation of dialogue systems, not only because they offer a good benchmark scenario to test and compare the systems on a common platform, but also because they often release the dialogue datasets for later evaluation.

Perhaps one of the most popular challenges is the Dialog State Tracking Challenge (DSTC),[18] which was previously mentioned in this section. DSTC was started in 2013 in order to provide a common testbed for the task of dialogue state tracking. It continued on a yearly basis with remarkable success. For its sixth edition, it was renamed as Dialog

[18] https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge/.

System Technology Challenges due to the interest of the research community in a wider variety of dialogue-related problems. Various well-known datasets have been produced and released for every edition: DSTC1 has human-computer dialogues in the bus timetable domain; DSTC2 and DSTC3 used human-computer dialogues in the restaurant information domain; DSTC4 dialogues were human-human and in the tourist information domain; DSTC5 also is from the tourist information domain, but training dialogues are provided in one language and test dialogues are in a different language. Finally, as the DSTC6 edition consisted of 3 parallel tracks, different datasets were released for each track, such as, a transaction dialogue dataset for the restaurant domain, two datasets that are part of OpenSubtitles and Twitter datasets, and different chat-oriented dialogue datasets with dialogue breakdown annotations in Japanese and English.

A more recent challenge that started in 2017 and continued into 2018, with its second edition being the Conversational Intelligence Challenge (ConvAI).[19] This challenge, conducted under the scope of NIPS, has the aim to unify the community around the task of building systems capable of intelligent conversations. In its first edition teams were expected to submit dialogue systems able to carry out intelligent and natural conversations about specific news articles with humans. The aim of the task of the second edition has been to model normal conversation when two interlocutors meet for the first time, and get to know each other. The dataset of this task consists of 10,981 dialogues with 164,356 utterances, and it is available in the ParlAI framework mentioned above.

Finally, the Alexa Prize[20] has attracted mass media and research attention alike. This annual competition for university teams is dedicated at accelerating the field of conversational AI in the framework of the Alexa technology. The participants have to create socialbots that can converse coherently and engagingly with humans on news events and popular topics such as entertainment, sports, politics, technology and fashion. Unfortunately, no datasets have been released.

## 7 Challenges and future trends

In the introduction, we stated that the goal of the dialogue evaluation is to find methods that are automated, repeatable, are correlated to human judgements, capable of differentiating among various dialogue strategies and explain which features of the dialogue system contribute to its quality. The main motivation behind this is the need to reduce the human evaluation effort as much as possible, since human involvement creates high costs and is highly time-consuming. In this survey, we presented the main concepts regarding evaluation of dialogue systems and showcased the most important methods. However, evaluation of dialogue systems is still an area of open research. In this section, we summarize the current challenges and future trends that we deem most important.

*Automation* The evaluation methods covered in this survey all achieve a certain degree of automation. However, the automation is achieved with significant engineering effort, or by loss of correlation to human judgements. Word-overlap metrics (see Sect. 4.3.1), which are borrowed from the machine translation and summarization community, are fully automated. However, they do not correlate with human judgements on the turn level. On the

[19] http://convai.io/.

[20] https://developer.amazon.com/alexaprize.

other hand, BLEU becomes more competitive when applied on the corpus-level or system-level (Galley et al. 2015; Lowe et al. 2017a). More recent metrics such as ΔBLEU and ADEM (see Sect. 4.3.1) have significantly higher correlations to human judgements while requiring a significant amount of human annotated data as well as thorough engineering.

Task-oriented dialogue systems can be evaluated semi-automatically or even fully automatically. These systems benefit from having a well-defined task, where success can be measured. Thus, user satisfaction modelling (see Sect. 3.4.1) as well as user simulations (see Sect. 3.4.2) exploit this to automate their evaluation. However, both approaches need a significant amount of engineering and human annotation: user satisfaction modelling usually requires prior annotation effort, which is followed by fitting a model that predicts the judgements. In addition to this effort, the process has to be potentially repeated for each new domain or new functionality that the dialogue system incorporates. Although in some cases the model fitted on the data for one dialogue system can be reused to predict another dialogue system, this is not always possible.

On the other hand, user simulations require two steps: gathering data to develop a first version of the simulation, and then building the actual user simulation. The first step is only required for user simulations that are based on training corpora (e.g. the neural user simulation). A significant drawback is that the user simulation is only capable of simulating the behaviour which is represented in the corpus or the rules. This means that it cannot cover unseen behaviour well. Furthermore, the user simulation can hardly be used to train or evaluate dialogue systems for other tasks or domains.

Automation is thus achieved to a certain degree, but with significant drawbacks. Hence, finding ways to facilitate the automation of evaluation methods is clearly an open challenge.

*High quality dialogues* One major objective for a dialogue system is to deliver high quality interactions with its users. However, it is often not clear how "high quality" is defined in this context or how to measure it. For task oriented dialogue systems, the mostly used definition of quality is often measured by means of task success and number of dialogue turns (e.g. a reward of 20 for task-success minus the number of turns needed to achieve the goal). However, this definition is not applicable to conversational dialogue systems and it might ignore other aspects of the interaction (e.g. frustration of the user). Thus, the current trend is to let humans judge the *appropriateness* of the system utterances. However, the notion of appropriateness is highly subjective and entails several finer-grained concepts (e.g. ability to maintain the topic, the coherence of the utterance, the grammatical correctness of the utterance itself, etc.). Currently, appropriateness is modelled by means of latent representations (e.g. ADEM), which are derived again from annotated data.

Other aspects of quality concern the purpose of the dialogue system in conjunction with the functionality of the system. For instance, Zhou et al. (2018) define the purpose of their conversational dialogue system to build an emotional bond between the dialogue system and the user. This goal differs significantly from the task of training a medical student in the interaction with patients. Both systems need to be evaluated with respect to their particular goal. The ability to build an emotional bond can be evaluated by means of the interaction length (longer interactions are an indicator of a higher user engagement), whereas training (or e-learning) systems are usually evaluated regarding their ability of selecting an appropriate utterance for the given context.

The target audience plays an important role as well. Since quality is mainly a subjective measure, different user groups prefer different types of interactions. For instance, depending on the level of domain knowledge, novice users prefer instructions that use less specialized wording, whereas domain experts might prefer a more specialized vocabulary.

The notion of quality is thus dependent on a large amount of factors. The evaluation needs to be adapted to take aspects such as the dialogue system's purpose, the target audience, and the dialogue system implementation itself into account.

*Lifelong learning* The notion of lifelong learning for machine learning systems has gained traction recently. The main concept of lifelong learning is that a deployed machine learning system continues to improve by interaction with its environment (Chen et al. 2016). Lifelong learning for dialogue systems is motivated by the fact that it is not possible to encounter all possible situations during training, thus, a component that allows the dialogue system to retrain itself and adapt its strategy during deployment seems the most logical solution.

The evaluation step is critical in order to achieve lifelong learning. Since the dialogue system relies on the ability to automatically find critical dialogue states where it needs assistance, a module is needed which is able to evaluate the ongoing dialogue. One step in this direction is done by Hancock et al. (2019), who present a solution that relies on a satisfaction module that is able of to classify the current dialogue state as either satisfactory or not. If this module finds an unsatisfactory dialogue state, a feedback module asks the user for feedback. The feedback data is then used to improve the dialogue system.

The aspect of lifelong learning brings a large variety of novel challenges. Firstly, the lifelong learning system requires a module that self-monitors its behaviour and notices when a dialogue is going wrong. For this, the module needs to rely on evaluation methods that work automatically, or at least semi-automatically. The second challenge lies in the evaluation of the lifelong learning system itself. The self-monitoring module as well as the adaptive behaviour need to be evaluated. This brings a new dimension of complexity into the evaluation procedure.

### 7.1 Conclusion

Evaluation is a critical task when developing and researching dialogue systems. Over the past decades, many methods and concepts have been proposed. These methods and concepts are related to the different requirements and functionalities of the dialogue systems. These are subsequently dependent on the current development stage of the dialogue system technology. Currently, the trend is moving towards building end-to-end trainable dialogue systems based on large amounts of data. These systems have different requirements for evaluation than a finite state, machine-based system. Thus, the problem of evaluation is evolving in tandem to the progress of the dialogue system technology itself. This survey presents the current state-of-the-art research in evaluation.

**Acknowledgements** We would like to thank Lina Scarborough for proofreading the manuscript. We would also like to thank our anonymous reviewers for their valuable reviews that helped improve the quality of this survey.

**Funding** This work was supported by the LIHLITH project in the framework of EU ERA-Net CHIST-ERA; the Swiss National Science Foundation [20CH21_174237]; the Spanish Research Agency [PCIN-2017-11, PCIN-2017-085/AEI]; Eneko Agirre and Arantxa Otegi received the support of the UPV/EHU [grant GIU16/16]; Agence Nationale pour la Recherche [ANR-17-CHR2-0001-03].

## Compliance with ethical standards

**Conflict of interest** There are no conflicts of interest to disclose.

## References

Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, Thoppilan R, Yang Z, Kulshreshtha A, Nemade G, Lu Y, et al. (2020) Towards a human-like open-domain chatbot. arXiv preprint arXiv:200109977

Ameixa D, Coheur L (2013) From subtitles to human interactions: introducing the SubTle Corpus. In: Technical report 2013

Austin JL (1962) How to do things with words. Oxford University Press, Oxford, William James

Banchs RE (2012) Movie-DiC: a Movie Dialogue Corpus for Research and Development. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, pp 203–207

Banchs RE, Li H (2012) IRIS: a chat-oriented dialogue system based on the vector space model. In: Proceedings of the ACL 2012 demonstrations, Jeju Island, Korea, pp 37–42

Bernardi R, Kirschner M (2010) From artificial questions to real user interaction logs: Real challenges for Interactive Question Answering systems. In: Proceedings of workshop on web logs and question answering (WLQA'10), Valletta, Malta

Black AW, Eskenazi M (2009) The Spoken Dialogue Challenge. In: Proceedings of the SIGDIAL 2009 conference: the 10th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, Stroudsburg, PA, USA, SIGDIAL '09, pp 337–340

Black AW, Burger S, Conkie A, Hastie H, Keizer S, Lemon O, Merigaud N, Parent G, Schubiner G, Thomson B, Williams JD, Yu K, Young S, Eskenazi M (2011) Spoken Dialog Challenge 2010: comparison of live and control test results. In: Proceedings of the SIGDIAL 2011 conference: The 12th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, Portland, Oregon, pp 2–7

Bordes A, Boureau YL, Weston J (2017) Learning end-to-end goal-oriented dialog. In: International conference on learning representations (ICLR) 2017, Toulon, France

Bowman SR, Vilnis L, Vinyals O, Dai A, Jozefowicz R, Bengio S (2016) Generating sentences from a continuous space. In: Proceedings of The 20th SIGNLL conference on computational natural language learning, Association for Computational Linguistics, Berlin, Germany, pp 10–21

Bruni E, Fernandez R (2017) Adversarial evaluation for open-domain dialogue generation. In: Proceedings of the SIGDIAL 2017 conference: The 18th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, pp 284–288

Budzianowski P, Wen TH, Tseng BH, Casanueva I, Stefan U, Osman R, Gašić M (2018) MultiWOZ: A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proceedings of the (2018) conference on empirical methods in natural language processing (EMNLP). Belgium, Brussels

Byrne B, Krishnamoorthi K, Sankar C, Neelakantan A, Goodrich B, Duckworth D, Yavuz S, Dubey A, Kim K, Cedilnik A (2019) Taskmaster-1: Toward a realistic and diverse dialog dataset. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, Association for Computational Linguistics, pp 4515–4524, https://doi.org/10.18653/v1/D19-1459

Campos JA, Otegi A, Soroa A, Deriu J, Cieliebak M, Agirre E (2019) Conversational QA for FAQs. In: 3rd Conversational AI: "Today's Practice and Tomorrow's Potential" workshop at NeurIPS 2019

Carletta J (1996) Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics 22(2):249–254

Charras F, Dubuisson Duplessis G, Letard V, Ligozat AL, Rosset S (2016) Comparing system-response retrieval models for open-domain and casual conversational agent. In: Workshop on Chatbots and Conversational Agent Technologies (WOCHAT)

Chen H, Liu X, Yin D, Tang J (2017) A Survey on dialogue systems: recent advances and new frontiers. Special interest group on knowledge discovery and data mining (SIGKDD) Explor Newsl 19(2):25–35

Chen Z, Liu B, Brachman R, Stone P, Rossi F (2016) Lifelong Machine Learning, 1st edn. Morgan & Claypool Publishers, San Rafael

Choi E, He H, Iyyer M, Yatskar M, Yih Wt, Choi Y, Liang P, Zettlemoyer L (2018) QuAC: Question answering in context. In: Proceedings of the (2018) conference on empirical methods in natural language processing (EMNLP). France, Paris

Chotimongkol A, Rudnicky AI (2001) N-best speech hypotheses reordering using linear regression. In: Dalsgaard P, Lindberg B, Benner H, Tan Z (eds) EUROSPEECH 2001 Scandinavia, 7th European conference on speech communication and technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3–7, 2001, ISCA, pp 1829–1832, http://www.isca-speech.org/archive/eurospeech_2001/e01_1829.html

Clark P, Etzioni O (2016) My computer is an honor student but how intelligent is it? standardized tests as a measure of ai. AI Mag 37(1):5–12. https://doi.org/10.1609/aimag.v37i1.2636

Colby KM (1981) Modeling a paranoid mind. Behav Brain Sci 4(4):515–534

Cole R (1999) Tools for research and education in speech science. In: Proceedings of the international conference of phonetic sciences, San Francisco, USA, pp 1277–1280

Collins E, Rozanov N, Zhang B (2019) LIDA: lightweight interactive dialogue annotator. In: Padó S, Huang R (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019—system demonstrations, Association for Computational Linguistics, pp 121–126, https://doi.org/10.18653/v1/D19-3021

Danescu C, Lee L (2011) Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In: Proceedings of the 2nd workshop on cognitive modeling and computational linguistics, Association for Computational Linguistics, pp 76–87

Dethlefs N, Hastie H, Cuayáhuitl H, Lemon O (2013) Conditional random fields for responsive surface realisation using global features. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, pp 1254–1263

DeVault D, Leuski A, Sagae K (2011) Toward learning and evaluation of dialogue policies with text examples. In: Proceedings of the SIGDIAL 2011 conference: the 12th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 39–48

Devlin J, Chang MW, Lee K, Toutanova K (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), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423

Diefenbach D, Lopez V, Singh K, Maret P (2018) Core techniques of question answering systems over knowledge bases: a survey. Knowl Inf Syst 55(3):529–569

Do P, Nguyen H, Tran C, Nguyen M, Nguyen M (2017) Legal question answering using ranking SVM and deep convolutional neural network. arXiv preprint arXiv:abs/1703.05320

Dubuisson DG, Letard V, Ligozat AL, Rosset S (2016) Purely corpus-based automatic conversation authoring. In: Proceedings of the tenth international conference on language resources and evaluation, European Language Resources Association (ELRA), Paris, France, LREC 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/396_Paper.pdf

Dubuisson DG, Charras F, Letard V, Ligozat AL, Rosset S (2017) Utterance retrieval based on recurrent surface text patterns. In: European conference on information retrieval, Aberdeen, Scotland UK, ECIR 2017, https://hal.archives-ouvertes.fr/hal-01436052/document

Dušek O, Jurcicek F (2016) Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, ACL 2016, pp 45–51

Dušek O, Novikova J, Rieser V (2020) Evaluating the state-of-the-art of end-to-end natural language generation: the E2E NLG challenge. Comput Speech Lang 59:123–156. https://doi.org/10.1016/j.csl.2019.06.009

Engel Y, Mannor S, Meir R (2005) Reinforcement learning with gaussian processes. In: Proceedings of the 22nd international conference on machine learning, ACM, Bonn, Germany, ICML '05, pp 201–208

Engelbrecht KP, Möller S, Schleicher R, Wechsung I (2008) Analysis of paradise models for individual users of a spoken dialog system. In: Electronic speech signal processing, proceedings of the 19th conference, Frankfurt am Main, Germany, ESSV 2008, pp 86–93, https://d-nb.info/990359174/04

Engelbrecht KP, Gödde F, Hartard F, Ketabdar H, Möller S (2009a) Modeling user satisfaction with Hidden Markov Model. In: Proceedings of the SIGDIAL 2009 conference: the 10th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, London, UK, SIGDIAL '09, pp 170–177, http://dl.acm.org/citation.cfm?id=1708376.1708402

Engelbrecht KP, Quade M, Möller S (2009b) Analysis of a new simulation approach to dialog system evaluation. Speech Commun 51(12):1234–1252, http://dx.doi.org/10.1016/j.specom.2009.06.007

Eric M, Krishnan L, Charette F, Manning CD (2017) Key-value retrieval networks for task-oriented dialogue. In: Proceedings of the SIGDIAL 2017 conference: the 18th annual meeting of the special interest group on discourse and dialogue, Saarbrücken, Germany, SIGDIAL'17, pp 37–49, https://doi.org/10.18653/v1/W17-5506, http://aclweb.org/anthology/W17-5506

Evanini K, Hunter P, Liscombe J, Suendermann D, Dayanidhi K, Pieraccini R (2008) Caller experience: a method for evaluating dialog systems and its automatic prediction. In: 2008 IEEE spoken language technology workshop, Goa, India, pp 129–132, https://doi.org/10.1109/SLT.2008.4777857

Fader A, Zettlemoyer L, Etzioni O (2013) Paraphrase-driven learning for open question answering. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, pp 1608–1618, https://www.aclweb.org/anthology/P13-1158

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382. https://doi.org/10.1037/h0031619

Furlanello T, Lipton ZC, Tschannen M, Itti L, Anandkumar A (2018) Born-again neural networks. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, PMLR, Proceedings of machine learning research, vol 80, pp 1602–1611, http://proceedings.mlr.press/v80/furlanello18a.html

Galley M, Brockett C, Sordoni A, Ji Y, Auli M, Quirk C, Mitchell M, Gao J, Dolan B (2015) deltaBLEU: a discriminative metric for generation tasks with intrinsically diverse targets. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (Volume 2: Short Papers), Association for Computational Linguistics, ACL 2015, pp 445–450, http://www.aclweb.org/anthology/P15-2073

Gandhe S, Traum D (2016) A Semi-automated Evaluation Metric for Dialogue Model Coherence, Springer International Publishing, Cham, pp 217–225. https://doi.org/10.1007/978-3-319-21834-2_19

Gandhe S, Traum DR (2013) Surface text based dialogue models for virtual humans. In: Proceedings of the SIGDIAL (2013) conference: the 14th annual meeting of the special interest group on discourse and dialogue. Metz, France, SIGDIAL, p 2013

Gandhe S, Whitman N, Traum D, Artstein R (2009) An integrated authoring tool for tactical questioning dialogue systems. In: 6th IJCAI Workshop on knowledge and reasoning in practical dialogue systems, Pasadena Conference Center, California, USA., pp 10–18

Gasic M, Breslin C, Henderson M, Kim D, Szummer M, Thomson B, Tsiakoulis P, Young S (2013) POMDP-based dialogue manager adaptation to extended domains. In: Proceedings of the SIGDIAL 2013 conference: the 14th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, Metz, France, SIGDIAL 2013, pp 214–222, http://www.aclweb.org/anthology/W13-4035

Gasic M, Kim D, Tsiakoulis P, Breslin C, Henderson M, Szummer M, Thomson B, Young SJ (2014) Incremental on-line adaptation of POMDP-based dialogue managers to extended domains. In: 15th annual conference of the international speech communication association, Singapore, INTERSPEECH 2014, pp 140–144, http://www.isca-speech.org/archive/interspeech_2014/i14_0140.html

Gašić M, Jurčíček F, Thomson B, Yu K, Young S (2011) On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In: 2011 IEEE workshop on automatic speech recognition understanding, pp 312–317, https://doi.org/10.1109/ASRU.2011.6163950

Ghazvininejad M, Brockett C, Chang MW, Dolan B, Gao J, Yih Wt, Galley M (2018) A knowledge-grounded neural conversation model. Thirty-second AAAI conference on artificial intelligence, New Orleans, Louisiana, USA, AAAI 2018:5110–5117

Godfrey JJ, Holliman EC, McDaniel J (1992) SWITCHBOARD: telephone speech corpus for research and development. In: [Proceedings] ICASSP-92: 1992 IEEE international conference on acoustics, speech, and signal processing, San Francisco, CA, USA, vol 1, pp 517–520, https://doi.org/10.1109/ICASSP.1992.225858

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27, NIPS 27, Curran Associates, Inc., pp 2672–2680, http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Gunasekara C, Kummerfeld JK, Polymenakos L, Lasecki WS (2019) DSTC7 Task 1: Noetic end-to-end response selection. In: 7th edition of the dialog system technology challenges at AAAI 2019, http://workshop.colips.org/dstc7/papers/dstc7_task1_final_report.pdf

Guo D, Tur G, Yih Wt, Zweig G (2014) Joint semantic utterance classification and slot filling with recursive neural networks. In: 2014 IEEE spoken language technology workshop (SLT), South Lake Tahoe, California, USA, IEEE 2014, pp 554–559, https://www.microsoft.com/en-us/research/wp-content/uploads/2014/12/SLT2014-daniel.pdf

Guo F, Metallinou A, Khatri C, Raju A, Venkatesh A, Ram A (2018) Topic-based evaluation for conversational bots. arXiv preprint arXiv:180103622

Gupta P, Mehri S, Zhao T, Pavel A, Eskenazi M, Bigham JP (2019) Investigating evaluation of open-domain dialogue systems with human generated multiple references. In: 20th annual meeting of the special interest group on discourse and dialogue

Hahn S, Dinarelli M, Raymond C, Lefèvre F, Lehen P, De Mori R, Moschitti A, Ney H, Riccardi G (2010) Comparing stochastic approaches to spoken language understanding in multiple languages. IEEE Trans Audio Speech Lang Process 16:1569–1583

Hancock B, Bordes A, Mazare PE, Weston J (2019) Learning from dialogue after deployment: feed yourself, Chatbot! In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, Florence, Italy, ACL 2019, pp 3667–3684, https://www.aclweb.org/anthology/P19-1358

Hara S (2010) Estimation method of user satisfaction using N-gram-based dialog history model for spoken dialog system. In: Proceedings of the seventh international conference on language resources and evaluation, Valletta, Malta, LREC'10, pp 78–83, http://www.lrec-conf.org/proceedings/lrec2010/pdf/579_Paper.pdf

Henderson M, Thomson B, Williams J (2013a) Dialog state tracking challenge 2 & 3. Technical report

Henderson M, Thomson B, Young S (2013b) Deep neural network approach for the dialog state tracking challenge. In: Proceedings of the SIGDIAL 2013 Conference: The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Metz, France, pp 467–471, http://www.aclweb.org/anthology/W13-4073

Henderson M, Thomson B, Williams J (2014) The Second Dialog State Tracking Challenge. In: Proceedings of the SIGDIAL 2014 Conference: The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Philadelphia, PA, USA, pp 263–272, https://www.microsoft.com/en-us/research/publication/the-second-dialog-state-tracking-challenge/

Higashinaka R, Minami Y, Dohsaka K (2010) Meguro T (2010) Issues in predicting user satisfaction transitions in dialogues: individual differences, evaluation criteria, and prediction models. In: Lee GG, Mariani J, Minker W, Nakamura S (eds) Second international workshop on spoken dialogue systems technology: spoken dialogue systems for ambient environments. Springer, Berlin Heidelberg, Gotemba, Shizuoka, Japan, WSDS, pp 48–60

Hirschman L, Dahl DA, McKay DP, Norton LM, Linebarger MC (1990) Beyond class A: a proposal for automatic evaluation of discourse. In: Proceedings of the speech and natural language workshop, Hidden Valley, Pennsylvania, USA, HLT, pp 109–113

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9:1735–1780

Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: Proceedings of the 34th international conference on machine learning, international convention centre, Sydney, Australia, ICML, pp 1587–1596, http://proceedings.mlr.press/v70/hu17e.html

Huang HY, Choi E, tau Yih W (2019) FlowQA: grasping flow in history for conversational machine comprehension. In: International conference on learning representations, https://openreview.net/forum?id=ByftGnR9KX

Iyyer M, Yih Wt, Chang MW (2017a) Search-based neural structured learning for sequential question answering. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, ACL, pp 1821–1831, https://doi.org/10.18653/v1/P17-1167, http://www.aclweb.org/anthology/P17-1167

Iyyer M, Yih Wt, Chang MW (2017b) Search-based neural structured learning for sequential question answering. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp 1821–1831, https://doi.org/10.18653/v1/P17-1167, https://www.aclweb.org/anthology/P17-1167

Joshi M, Choi E, Weld D, Zettlemoyer L (2017) TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp 1601–1611, https://doi.org/10.18653/v1/P17-1147, https://www.aclweb.org/anthology/P17-1147

Ju Y, Zhao F, Chen S, Zheng B, Yang X, Liu Y (2019) Technical report on conversational question answering

Jurafsky D, Martin JH (2017) Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 3rd edn. Prentice Hall PTR, USA

Jurcícek F, Keizer S, Gasic M, Mairesse F, Thomson B, Yu K, Young SJ (2011) Real user evaluation of spoken dialogue systems using amazon mechanical turk. 12th annual conference of the international speech communication association. Florence, Italy, INTERSPEECH, pp 3061–3064

Kannan A, Vinyals O (2016) Adversarial evaluation of dialogue models. In: Workshop on adversarial training at neural information processing systems 2016

Kelly D, Kantor PB, Morse EL, Scholtz J, Sun Y (2009) Questionnaires for eliciting evaluation data from users of interactive question answering systems. Nat Lang Eng 15(1):119–141

Kenny PG, Parsons TD, Rizzo AA (2009) Human computer interaction in virtual standardized patient systems. In: Proceedings of the 13th international conference on human-computer interaction. Part IV: interacting in various application domains, Springer-Verlag, Berlin, Heidelberg, pp 514–523, http://dx.doi.org/10.1007/978-3-642-02583-9_56

Kim S, D'Haro LF, Banchs RE, Williams JD, Henderson M, Yoshino K (2016) The fifth dialog state tracking challenge. In: 2016 IEEE Spoken Language Technology Workshop (SLT), pp 511–517, https://doi.org/10.1109/SLT.2016.7846311

Kočiský T, Schwarz J, Blunsom P, Dyer C, Hermann KM, Melis G, Grefenstette E (2018) The narrativeQA reading comprehension challenge. Trans Assoc Computational Ling 6:317–328. https://doi.org/10.1162/tacl_a_00023

Kolomiyets O, Moens MF (2011) A Survey on Question Answering Technology from an Information Retrieval Perspective. Inf Sci 181(24):5412–5434. https://doi.org/10.1016/j.ins.2011.07.047

Konstantinova N, Orasan C (2013) Interactive Question Answering. In: Emerging applications of natural language processing: concepts and new research, pp 149–169

Kreyssig F, Casanueva I, Budzianowski P, Gasic M (2018) Neural user simulation for corpus-based policy optimisation for spoken dialogue systems. arXiv preprint arXiv:1805.06966

Lai G, Xie Q, Liu H, Yang Y, Hovy E (2017) RACE: large-scale ReAding comprehension dataset from examinations. In: Proceedings EMNLP 2017—conference on empirical methods in natural language processing, pp 785–794, arXiv:1704.04683

Lamel L, Rosset S, Gauvain JL, Bennacef S, Garnier-Rizet M, Prouts B (2000) The limsi arise system. Speech Commun 31(4):339–353

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942

Larson S, Mahendran A, Peper JJ, Clarke C, Lee A, Hill P, Kummerfeld JK, Leach K, Laurenzano MA, Tang L, Mars J (2019) An evaluation dataset for intent classification and out-of-scope prediction. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 1311–1316, https://doi.org/10.18653/v1/D19-1131, https://www.aclweb.org/anthology/D19-1131

Lavie A, Denkowski MJ (2009) The meteor metric for automatic evaluation of machine translation. Mach Transl 23(2-3):105–115, http://dx.doi.org/10.1007/s10590-009-9059-4

Lee C, Jung S, Kim S, Lee GG (2009) Example-based dialog modeling for practical multi-domain dialog system. Speech Commun 51(5):466–484

Lee S, Schulz H, Atkinson A, Gao J, Suleman K, El Asri L, Adada M, Huang M, Sharma S, Tay W, Li X (2019) Multi-domain task-completion dialog challenge. In: Dialog system technology challenges 8

Leech GN (1993) 100 million words of english: the british national corpus (BNC). English Today 28:9–15. https://doi.org/10.1017/S0266078400006854

Lemon O, Pietquin O (2012) Data-driven methods for adaptive spoken dialogue systems: computational learning for conversational interfaces. Springer, Berlin

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys Doklady 10(8):707–710

Levin E, Pieraccini R, Eckert W (1998) Using Markov decision process for learning dialogue strategies. In: Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, Seattle, WA, USA, ICASSP, vol 1, pp 201–204, https://doi.org/10.1109/ICASSP.1998.674402

Li H, Min MR, Ge Y, Kadav A (2017a) A context-aware attention network for interactive question answering. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '17, pp 927–935, http://doi.acm.org/10.1145/3097983.3098115

Li J, Galley M, Brockett C, Gao J, Dolan B (2016a) A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, Association for Computational Linguistics, San Diego, California, pp 110–119, http://www.aclweb.org/anthology/N16-1014

Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J (2016b) Deep reinforcement learning for dialogue generation. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Association for Computational Linguistics, Austin, Texas, EMNLP '16, pp 1192–1202, https://doi.org/10.18653/v1/D16-1127, http://www.aclweb.org/anthology/D16-1127

Li X, Chen YN, Li L, Gao J, Celikyilmaz A (2017b) End-to-end task-completion neural dialogue systems. In: Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, IJCNLP, pp 733–743, http://aclweb.org/anthology/I17-1074

Li Y, Su H, Shen X, Li W, Cao Z, Niu S (2017c) DailyDialog: A manually labelled multi-turn dialogue dataset. In: Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, pp 986–995, https://www.aclweb.org/anthology/I17-1099

Lin CY (2004) ROUGE: a package for automatic evaluation of summaries. In: Marie-Francine Moens SS (ed) Text summarization branches out: proceedings of the ACL-04 workshop, Association for Computational Linguistics, Barcelona, Spain, pp 74–81, http://www.aclweb.org/anthology/W04-1013

Liu B, Tür G, Hakkani-Tür D, Shah P, Heck L (2018) Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, USA, NAACL-HLT '18, pp 2060–2069, http://aclweb.org/anthology/N18-1187

Liu CW, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J (2016) How NOT To evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Association for Computational Linguistics, Austin, Texas, pp 2122–2132, https://doi.org/10.18653/v1/D16-1230, http://www.aclweb.org/anthology/D16-1230

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J (2016) On the evaluation of dialogue systems with next utterance classification. In: Proceedings of the SIGDIAL 2016 conference: the 17th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, Los Angeles, CA, USA, pp 264–269, http://www.aclweb.org/anthology/W16-3634

Lowe R, Noseworthy M, Serban IV, Angelard-Gontier N, Bengio Y, Pineau J (2017a) Towards an automatic turing test: learning to evaluate dialogue responses. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, ACL '17, pp 1116–1126, https://doi.org/10.18653/v1/P17-1103, http://www.aclweb.org/anthology/P17-1103

Lowe R, Pow N, Serban IV, Charlin L, Liu CW, Pineau J (2017b) Training end-to-end dialogue systems with the ubuntu dialogue corpus. Dialogue Discourse 8(1):31–65

Lowe RJ, Pow N, Serban I, Pineau J (2015) The Ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of the SIGDIAL 2015 conference: the 16th annual meeting of the special interest group on discourse and dialogue, Association for Computational Linguistics, Prague, Czech Republic, pp 285–294, http://aclweb.org/anthology/W15-4640

Lu X (2012) The relationship of lexical richness to the quality of ESL learners' oral narratives. Modern Lang J 96(2):190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Mairesse F, Gašić M, Jurčíček F, Keizer S, Thomson B, Yu K, Young S (2010) Phrase-based statistical language generation using graphical models and active learning. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics, Uppsala, Sweden, ACL '10, pp 1552–1561, https://www.aclweb.org/anthology/P10-1157

Mazza R, Ambrosini L, Catenazzi N, Vanini S, Tuggener D, Tavarnesi G (2018) Behavioural simulator for professional training based on natural language interaction. In: 10th international conference on education and new learning technologies, Palma, Mallorca, Spain, EDULEARN18, pp 3204–3214, http://repository.supsi.ch/9776/1/edulearn18-paper-lifelike.pdf

McTear M, O'Neill I, Hanna P, Liu X (2005) Handling errors and determining confirmation strategies–an object-based approach. Speech Commun 45(3):249–269

Mei H, Bansal M, Walter MR (2016) What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. In: Proceedings of the 2016 conference of the North American Chapter of

the Association for Computational Linguistics: human language technologies, San Diego, California, NAACL-HLT, pp 720–730, https://www.aclweb.org/anthology/N16-1086

Mesnil G, Dauphin Y, Yao K, Bengio Y, Deng L, Hakkani-Tur D, He X, Heck L, Tur G, Yu D, Zweig G (2015) Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Trans Audio, Speech Lang Process 23(3):530–539

Metallinou A, Bohus D, Williams J (2013) Discriminative state tracking for spoken dialog systems. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, pp 466–475, http://www.aclweb.org/anthology/P13-1046

Miller A, Feng W, Batra D, Bordes A, Fisch A, Lu J, Parikh D, Weston J (2017) ParlAI: a dialog research software platform. In: Proceedings of the 2017 conference on empirical methods in natural language processing: system demonstrations, EMNLP '17, pp 79–84, https://www.aclweb.org/anthology/D17-2014

Mishra A, Jain SK (2016) A survey on question answering systems with classification. J King Saud Univ Comput Inf Sci 28(3):345–361. https://doi.org/10.1016/j.jksuci.2014.10.007

Möller S, Krebber J, Raake A, Smeele P, Rajman M, Melichar M, Pallotta V, Tsakou G, Kladis B, Vovos A, Hoonhout J, Schuchardt D, Fakotakis N, Ganchev T, Potamitis I (2004) INSPIRE: evaluation of a smart-home system for infotainment management and device control. In: Proceedings of the fourth international conference on language resources and evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, http://www.lrec-conf.org/proceedings/lrec2004/pdf/12.pdf

Möller S, Englert R, Engelbrecht K, Hafner V, Jameson A, Oulasvirta A, Raake A, Reithinger N (2006) MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In: Ninth international conference on spoken language processing, INTERSPEECH—ICSLP 2006, pp 1786–1789, https://www.isca-speech.org/archive/interspeech_2006/i06_1131.html

Mrkšić N, Ó Séaghdha D, Wen TH, Thomson B, Young S (2017) Neural belief tracker: data-driven dialogue state tracking. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, ACL '17, pp 1777–1788, https://doi.org/10.18653/v1/P17-1163, http://aclweb.org/anthology/P17-1163

Novikova J, Dušek O, Rieser V (2017) The E2E dataset: new challenges for end-to-end generation. In: Proceedings of the 18th annual meeting of the special interest group on discourse and dialogue, Saarbrücken, Germany, SIGDIAL '17, pp 201–206, https://www.aclweb.org/anthology/W17-5525, arXiv:1706.09254

Paek T (2006) Reinforcement learning for spoken dialogue systems: comparing strengths and weaknesses for practical deployment. In: Proceedings of dialog-on-dialog workshop, interspeech, Pittsburgh, PA, USA, http://www.ling.helsinki.fi/~kjokinen/ICSLP06-DoD/Programme/PaekTim.pdf

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, ACL '02, pp 311–318, http://www.aclweb.org/anthology/P02-1040

Peñas A, Magnini B, Forner P, Sutcliffe R, Rodrigo Á, Giampiccolo D (2012) Question answering at the cross-language evaluation forum 2003–2010. Lang Resour Evaluat 46(2):177–217. https://doi.org/10.1007/s10579-012-9177-0

Perez J, Boureau YL, Bordes A (2017) Dialog system and technology challenge 6 overview of track 1 - end-to-end goal-oriented dialog learning. Technical report

Peskov D, Clarke N, Krone J, Fodor B, Zhang Y, Youssef A, Diab M (2019) Multi-domain goal-oriented dialogues (MultiDoGO): strategies toward curating and annotating large scale dialogue data. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 4526–4536, https://doi.org/10.18653/v1/D19-1460, https://www.aclweb.org/anthology/D19-1460

Pietquin O, Hastie H (2013) A survey on metrics for the evaluation of user simulations. Knowl Eng Rev 28(1):59–73. https://doi.org/10.1017/S0269888912000343

Powers DMW (2012) The Problem with Kappa. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics, Avignon, France, EACL '13, pp 345–355, http://www.aclweb.org/anthology/E12-1035

Qu C, Yang L, Croft WB, Trippas JR, Zhang Y, Qiu M (2018) Analyzing and characterizing user intent in information-seeking conversations. In: The 41st international ACM SIGIR conference on research & development in information retrieval, Ann Arbor, MI, USA, SIGIR 2018, pp 989–992, https://doi.org/10.1145/3209978.3210124

Qu C, Yang L, Qiu M, Zhang Y, Chen C, Croft WB, Iyyer M (2019) Attentive history selection for conversational question answering. In: Proceedings of the 28th ACM international conference on information and knowledge management, Association for Computing Machinery, New York, NY, USA, CIKM '19, pp 1391–1400, https://doi.org/10.1145/3357384.3357905,

Qu Y, Green N (2002) A constraint-based approach for cooperative information-seeking dialogue. In: Proceedings of the international natural language generation conference, Harriman, New York, USA, INLG, pp 136–143

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Association for Computational Linguistics, Austin, Texas, pp 2383–2392, https://doi.org/10.18653/v1/D16-1264, https://www.aclweb.org/anthology/D16-1264

Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, pp 784–789, https://doi.org/10.18653/v1/P18-2124, https://www.aclweb.org/anthology/P18-2124,

Rambow O, Bangalore S, Walker M (2001) Natural language generation in dialog systems. In: Proceedings of the first international conference on Human language technology (HLT) research, San Diego, USA, pp 67–73

Rastogi A, Zang X, Sunkara S, Gupta R, Khaitan P (2019) Towards scalable multi-domain conversational agents: the schema-guided dialogue dataset. arXiv preprint arXiv:1909.05855

Reddy S, Chen D, Manning CD (2018) CoQA: a conversational question answering challenge. Trans Assoc Comput Linguist 7:249–266

Richardson M, Burges CJ, Renshaw E (2013) MCTest: a challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Association for Computational Linguistics, Seattle, Washington, USA, pp 193–203, https://www.aclweb.org/anthology/D13-1020

Rieser V, Lemon O (2009) Does this list contain what you were searching for? Learning adaptive dialogue strategies for interactive question answering. Nat Lang Eng 15(1):55–72. https://doi.org/10.1017/S1351324908004907

Ritter A, Cherry C, Dolan B (2010) Unsupervised modeling of twitter conversations. In: Human language technologies: the 2010 annual conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '10, pp 172–180, http://dl.acm.org/citation.cfm?id=1857999.1858019

Ritter A, Cherry C, Dolan WB (2011) Data-driven response generation in social media. In: Proceedings of the conference on empirical methods in natural language processing, Edinburgh, Scotland, UK., EMNLP '11, pp 583–593, http://dl.acm.org/citation.cfm?id=2145432.2145500

Rodrigo A, Peñas A, Miyao Y, Kando N (2018) Do systems pass university entrance exams? Inf Process Manag 54(4):564–575. https://doi.org/10.1016/J.IPM.2018.03.002

Rogers A, Kovaleva O, Downey M, Rumshisky A (2020a) Getting closer to AI complete question answering: a set of prerequisite real tasks. In Proceedings of the AAAI conference on artificial intelligence

Rogers A, Kovaleva O, Rumshisky A (2020b) A primer in BERTology: What we know about how BERT works arXiv:2002.12327

Saha A, Pahuja V, Khapra MM, Sankaranarayanan K, Chandar S (2018) Complex sequential question answering: towards learning to converse over linked question answer pairs with a knowledge graph. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, pp 705–713, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17181

Sai AB, Gupta MD, Khapra MM, Srinivasan M (2019) Re-evaluating adem: a deeper look at scoring dialogue responses. In: Proceedings of the thirty-third AAAI conference on artificial intelligence, Honolulu, Hawaii, USA, AAAI'19, vol 33, pp 6220–6227, https://aaai.org/ojs/index.php/AAAI/article/view/4581

Sarrouti M, Ouatik El Alaoui S (2017) A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. J Biomed Inf 68(C):96–103. https://doi.org/10.1016/j.jbi.2017.03.001

Schatzmann J, Weilhammer K, Stuttle M, Young S (2006) A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowl Eng Rev 21(2):97–126

Schatzmann J, Thomson B, Weilhammer K, Ye H, Young S (2007) Agenda-based user simulation for bootstrapping a POMDP dialogue system. In: Human language technologies 2007: the conference

of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, Rochester, New York, NAACL-Short '07, pp 149–152, http://dl.acm.org/citation.cfm?id=1614108.1614146

Schatztnann J, Stuttle MN, Weilhammer K, Young S (2005) Effects of the user model on simulation-based learning of dialogue strategies. In: IEEE workshop on automatic speech recognition and understanding, San Juan, Puerto Rico, ASRU, pp 220–225, https://ieeexplore.ieee.org/document/1566539

Schmitt A, Ultes S (2015) Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts–and how it relates to user satisfaction. Speech Commun 74:12–36

Schmitt A, Ultes S, Minker W (2012) A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In: Chair) NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the eight international conference on language resources and evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey

Schrading JN (2015) Analyzing domestic abuse using natural language processing on social media data. Master's thesis, Rochester Institute of Technology, http://scholarworks.rit.edu/theses

Searle JR (1969) Speech acts: an essay in the philosophy of language. Cambridge University Press, Cambridge

Searle JR (1975) Indirect speech acts. In: Cole P, Morgan J (eds) Syntax and semantics 3: speech acts. Academic Press, New York, pp 59–82

Semeniuta S, Severyn A, Barth E (2017) A hybrid convolutional variational autoencoder for text generation. In: Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, Denmark, EMNLP, pp 627–637, https://www.aclweb.org/anthology/D17-1066

Serban IV, Sordoni A, Bengio Y, Courville A, Pineau J (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, AAAI Press, Phoenix, Arizona, USA, AAAI'16, pp 3776–3783, http://dl.acm.org/citation.cfm?id=3016387.3016435

Serban IV, Klinger T, Tesauro G, Talamadupula K, Zhou B, Bengio Y, Courville AC (2017a) Multiresolution recurrent neural networks: an application to dialogue response generation. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, San Francisco, California, USA, AAAI '17, pp 3288–3294, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14571

Serban IV, Sankar C, Germain M, Zhang S, Lin Z, Subramanian S, Kim T, Pieper M, Chandar S, Ke NR, et al. (2017b) A deep reinforcement learning chatbot. arXiv preprint arXiv:1709.02349

Serban IV, Sordoni A, Lowe R, Charlin L, Pineau J, Courville A, Bengio Y (2017c) A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the thirty-first aaai conference on artificial intelligence, San Francisco, California USA, AAAI'17, pp 3295–3301, https://dl.acm.org/doi/10.5555/3298023.3298047

Serban IV, Lowe R, Henderson P, Charlin L, Pineau J (2018) A survey of available corpora for building data-driven dialogue systems: the journal version. Dialogue Discourse 1(9):1–49

Shang L, Lu Z, Li H (2015) Neural responding machine for short-text conversation. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long Papers), Beijing, China, ACL - IJCNLP '15, pp 1577–1586, http://www.aclweb.org/anthology/P15-1152

Singh SP, Kearns MJ, Litman DJ, Walker MA (2000) Reinforcement learning for spoken dialogue systems. In: Solla SA, Leen TK, Müller K (eds) Advances in neural information processing systems 12, MIT Press, pp 956–962, http://papers.nips.cc/paper/1775-reinforcement-learning-for-spoken-dialogue-systems.pdf

Sordoni A, Galley M, Auli M, Brockett C, Ji Y, Mitchell M, Nie JY, Gao J, Dolan B (2015) 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, Beijing, China, ACL—IJCNLP '15, pp 196–205, https://doi.org/10.3115/v1/N15-1020, http://www.aclweb.org/anthology/N15-1020

Stent A, Prasad R, Walker M (2004) Trainable sentence planning for complex information presentation in spoken dialog systems. In: Proceedings of the 42nd annual meeting of the Association for Computational Linguistics, Barcelona, Spain, ACL '04, pp 79–86, https://www.aclweb.org/anthology/P04-1011

Sugiyama H, Meguro T, Higashinaka R (2019) Automatic evaluation of chat-oriented dialogue systems using large-scale multi-references, Springer International Publishing, Cham, pp 15–25. https://doi.org/10.1007/978-3-319-92108-2_2,

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the 27th international conference on neural information processing systems—Volume 2, MIT

Press, Cambridge, MA, USA, NIPS'14, pp 3104–3112, http://dl.acm.org/citation.cfm?id=2969033.2969173

Talmor A, Berant J (2018) The web as a knowledge-base for answering complex questions. In: Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, pp 641–651, https://doi.org/10.18653/v1/N18-1059, https://www.aclweb.org/anthology/N18-1059

Tao C, Mou L, Zhao D, Yan R (2018) Ruber: an unsupervised method for automatic evaluation of open-domain dialog systems. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16179/15752

Tiedemann J (2009) News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In: Recent advances in natural language processing, vol 5, pp 237–248

Tiedemann J (2012) Parallel Data, Tools and Interfaces in OPUS. In: Chair) NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the eight international conference on language resources and evaluation (LREC'12), European Language Resources Association (ELRA)

Traum DR (1999) Speech acts for dialogue agents, Springer Netherlands, Dordrecht, pp 169–201. https://doi.org/10.1007/978-94-015-9204-8_8

Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K (2017) NewsQA: a machine comprehension dataset. In: Proceedings of the 2nd workshop on representation learning for NLP, Association for Computational Linguistics, Vancouver, Canada, pp 191–200, https://doi.org/10.18653/v1/W17-2623, https://www.aclweb.org/anthology/W17-2623

Tur G, De Mori R (2011) Spoken language understanding: systems for extracting semantic information from speech. Wiley, Hoboken

Tur G, Mori RD (2011) Spoken language understanding: systems for extracting semantic information from speech. Wiley, Hoboken

Turing AM (1950) Computing machinery and intelligence. Mind LIX(236):433–460. https://doi.org/10.1093/mind/LIX.236.433

Ultes S, Schmitt A, Minker W (2013) On quality ratings for spoken dialogue systems–experts vs. users. In: Proceedings of the 2013 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, Atlanta, Georgia, USA, NAACL—HLT'13, pp 569–578, https://www.aclweb.org/anthology/N13-1064

Ultes S, Rojas Barahona LM, Su PH, Vandyke D, Kim D, Casanueva In, Budzianowski P, Mrkšić N, Wen TH, Gasic M, Young S (2017) PyDial: a multi-domain statistical dialogue system toolkit. In: Proceedings of ACL 2017, System Demonstrations, Vancouver, Canada, pp 73–78

van Schooten B, Rosset S, Galibert O, Max A, op den Akker R, Illouz G (2007) Handling speech input in the Ritel QA dialogue system. In: 8th annual conference of the international speech communication Association, Antwerp, Belgium, INTERSPEECH 2007, pp 126–129, https://www.isca-speech.org/archive/interspeech_2007/i07_0126.html

Vinyals O, Le Q (2015) A neural conversational model. arXiv preprint arXiv:150605869

Voorhees EM (2006) Evaluating question answering system performance, Springer Netherlands, Dordrecht, pp 409–430. https://doi.org/10.1007/978-1-4020-4746-6_13

Walker MA, Litman DJ, Kamm CA, Abella A (1997) PARADISE: a framework for evaluating spoken dialogue agents. In: Proceedings of the Eighth Conference on European chapter of the association for computational linguistics, Madrid, Spain, EACL '97, pp 271–280, https://doi.org/10.3115/979617.979652

Walker MA, Kamm CA, Litman DJ (2000) Towards developing general models of usability with PARADISE. Nat Lang Eng 6(3–4):363–377. https://doi.org/10.1017/S1351324900002503

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S (2018) GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, Association for Computational Linguistics, Brussels, Belgium, pp 353–355, https://doi.org/10.18653/v1/W18-5446, https://www.aclweb.org/anthology/W18-5446

Wang Z, Wen TH, Su PH, Stylianou Y (2015) Learning domain-independent dialogue policies via ontology parameterisation. In: Proceedings of the SIGDIAL 2015 conference: the 16th annual meeting of the special interest group on discourse and dialogue, Prague, Czech Republic, SIGDIAL '15, pp 412–416, https://doi.org/10.18653/v1/W15-4654, http://www.aclweb.org/anthology/W15-4654

Weizenbaum J (1966) ELIZA–a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45. https://doi.org/10.1145/365153.365168

Wen TH, Gašić M, Mrkšić N, Su PH, Vandyke D, Young S (2015) Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, EMNLP '15

Wen TH, Gašić M, Mrkšić N, Rojas-Barahona LM, Su PH, Vandyke D, Young S (2016) Multi-domain neural network language generation for spoken dialogue systems. In: Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, san Diego, California, NAACL -HLT '16, pp 120–129

Wen TH, Vandyke D, Mrkšić N, Gasic M, Rojas Barahona LM, Su PH, Ultes S, Young S (2017) A network-based end-to-end trainable task-oriented dialogue system. In: Proceedings of the 15th conference of the european chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, EACL '17, pp 438–449, http://aclweb.org/anthology/E17-1042

Williams J, Raux A, Ramachandran D, Black A (2013) The dialog state tracking challenge. In: Proceedings of the SIGDIAL 2013 conference, Association for Computational Linguistics, Metz, France, pp 404–413

Williams J, Raux A, Henderson M (2016) The dialog state tracking challenge series: a review. Dialogue & Discourse https://www.microsoft.com/en-us/research/publication/the-dialog-state-tracking-challenge-series-a-review/

Xing C, Wu W, Wu Y, Liu J, Huang Y, Zhou M, Ma W (2017) Topic aware neural response generation. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, San Francisco, California, USA, AAAI '17, pp 3351–3357, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563

Yang Y, Yih Wt, Meek C (2015) WikiQA: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on empirical methods in natural language processing, Association for Computational Linguistics, Lisbon, Portugal, pp 2013–2018, https://doi.org/10.18653/v1/D15-1237, https://www.aclweb.org/anthology/D15-1237

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems, pp 5754–5764

Yao K, Peng B, Zhang Y, Yu D, Zweig G, Shi Y (2014) Spoken language understanding using long short-term memory neural networks. In: Spoken language technology workshop (SLT), IEEE, South Lake Tahoe, NV, USA, IEEE 2014, pp 189–194, https://doi.org/10.1109/SLT.2014.7078572, https://ieeexplore.ieee.org/document/7078572

Yeh YT, Chen YN (2019) FlowDelta: modeling flow information gain in reasoning for conversational machine comprehension. In: Proceedings of the 2nd workshop on machine reading for question answering, Association for Computational Linguistics, Hong Kong, China, pp 86–90, https://doi.org/10.18653/v1/D19-5812, https://www.aclweb.org/anthology/D19-5812

Young S (2007) CUED standard dialogue acts. Report, Cambridge University, Engineering Department http://mi.eng.cam.ac.uk/research/dialogue/LocalDocs/dastd.pdf

Young S, Schatzmann J, Weilhammer K, Ye H (2007) The hidden information state approach to dialog management. In: IEEE International conference on acoustics, speech and signal processing, Honolulu, HI, USA, ICASSP '07, vol 4, pp 149–152, http://svr-ftp.eng.cam.ac.uk/~sjy/papers/yswy07.pdf

Young S, Gašić M, Keizer S, Mairesse F, Schatzmann J, Thomson B, Yu K (2010) The hidden information state model: a practical framework for POMDP-based spoken dialogue management. Comput Speech Lang 24(2):150–174. https://doi.org/10.1016/j.csl.2009.04.001

Young S, Gašić M, Thomson B, Williams JD (2013) POMDP-based statistical spoken dialog systems: a review. Proc IEEE 101(5):1160–1179. https://doi.org/10.1109/JPROC.2012.2225812

Zhang X, Wang H (2016) A joint model of intent determination and slot filling for spoken language understanding. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, New York, New York, USA, IJCAI'16, pp 2993–2999, https://www.ijcai.org/Proceedings/16/Papers/425.pdf

Zhao T, Eskenazi M (2016) Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In: Proceedings of the SIGDIAL 2016 conference: the 17th annual meeting of the special interest group on discourse and dialogue, Los Angeles, CA, USA, SIGDIAL'16, pp 1–10, https://doi.org/10.18653/v1/W16-3601, http://www.aclweb.org/anthology/W16-3601

Zhao T, Zhao R, Eskenazi M (2017) Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp 654–664, https://doi.org/10.18653/v1/P17-1061, https://www.aclweb.org/anthology/P17-1061

Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing Twitter and traditional media using topic models. In: Proceedings of the 33rd European conference on advances in information

retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'11, pp 338–349, http://dl.acm.org/citation.cfm?id=1996889.1996934

Zhou L, Gao J, Li D, Shum HY (2018) The Design and implementation of XiaoIce, an empathetic social chatbot. arXiv preprint arXiv:1812.08989