Dialogue & Discourse 9(1) (2018) 1–49 doi: 10.5087/dad.2018.101

A Survey of Available Corpora for Building Data-Driven Dialogue Systems: The Journal Version

Iulian Vlad Serban {IULIAN.VLAD.SERBAN} AT UMONTREAL DOT CA
MILA, DIRO, Université de Montréal
2920 chemin de la Tour, Montréal, QC H3C 3J7, Canada

Ryan Lowe {RYAN.LOWE} AT MAIL DOT MCGILL DOT CA
School of Computer Science, McGill University
3480 University st, Montréal, QC H3A 0E9, Canada

Peter Henderson {PETER.HENDERSON} AT MAIL DOT MCGILL DOT CA
School of Computer Science, McGill University
3480 University st, Montréal, QC H3A 0E9, Canada

Laurent Charlin* {LAURENT.CHARLIN} AT HEC DOT CA
Department of Decision Sciences, HEC Montréal
3000, chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7, Canada

Joelle Pineau {JPINEAU} AT CS DOT MCGILL DOT CA
School of Computer Science, McGill University
3480 University st, Montréal, QC H3A 0E9, Canada

Editor: David Traum

Submitted 12/2015; Accepted 03/2018; Published online 05/2018

Abstract

During the past decade, several areas of speech and language understanding have witnessed substantial breakthroughs from the use of data-driven models. In the area of dialogue systems, the trend is less obvious, and most practical systems are still built through significant engineering and expert knowledge. Nevertheless, several recent results suggest that data-driven approaches are feasible and quite promising. To facilitate research in this area, we have carried out a wide survey of publicly available datasets suitable for data-driven learning of dialogue systems. We discuss important characteristics of these datasets, how they can be used to learn various components of a dialogue system, and their other potential uses. We also examine methods for transfer learning between datasets and the use of external knowledge. Finally, we discuss appropriate choices of evaluation metrics for the learning objective.

∗. Work done primarily while author was at McGill University

1. Introduction

Dialogue systems, also known as interactive conversational agents, virtual agents or sometimes chatterbots, are useful in a wide range of applications ranging from technical support services to language learning tools and entertainment (Young et al., 2013; Shawar and Atwell, 2007b). Large-scale data-driven methods, which use recorded data to automatically infer knowledge and strategies, are becoming increasingly important in speech and language understanding and generation. Speech recognition performance has increased tremendously over the last decade due to innovations in deep learning architectures (Hinton et al., 2012; Goodfellow et al., 2015). Similarly, a wide range of data-driven machine learning methods have been shown to be effective for natural language processing, including tasks relevant to dialogue, such as dialogue act classification (Reithinger and Klesen, 1997; Stolcke et al., 2000), dialogue state tracking (Thomson and Young, 2010; Wang and Lemon, 2013; Ren et al., 2013; Henderson et al., 2013; Williams et al., 2013; Henderson et al., 2014c; Kim et al., 2015), natural language generation (Langkilde and Knight, 1998; Oh and Rudnicky, 2000; Walker et al., 2002; Ratnaparkhi, 2002; Stent et al., 2004; Rieser and Lemon, 2010; Mairesse et al., 2010; Mairesse and Young, 2014; Wen et al., 2015a; Sharma et al., 2016), and dialogue policy learning (Young et al., 2013). Recent advances in computing power, the availability of large public datasets, and the development of efficient and effective machine learning models has led to increases attention and success in data-driven dialogue systems.

Importantly, the use of machine learning methods — such as neural networks — require an understanding of the availability, requirements, and uses of available dialogue corpora. To this end, this paper presents a broad survey of available dialogue corpora.

Corpus-based learning is not the only approach to training dialogue systems. Researchers have also proposed training dialogue systems online through live interaction with humans, and offline using user simulator models and reinforcement learning methods (Levin et al., 1997; Georgila et al., 2006; Paek, 2006; Schatzmann et al., 2007; Jung et al., 2009; Schatzmann and Young, 2009; Gašić et al., 2010, 2011; Daubigney et al., 2012; Gašić et al., 2012; Su et al., 2013; Gasic et al., 2013; Pietquin and Hastie, 2013; Young et al., 2013; Mohan and Laird, 2014; Su et al., 2015; Piot et al., 2015; Cuayáhuitl et al., 2015; Hiraoka et al., 2016; Fatemi et al., 2016; Asri et al., 2016; Williams and Zweig, 2016; Su et al., 2016). However, these approaches are beyond the scope of this survey unless the simulators are built from dialogue corpora.

This survey is structured as follows. In the next section, we give a high-level overview of dialogue systems. We briefly discuss the purpose and goal of dialogue systems. Then we describe the individual system components that are relevant for data-driven approaches as well as holistic end-to-end dialogue systems. In Section 3, we discuss types of dialogue interactions and aspects relevant to building data-driven dialogue systems, from a corpus perspective, as well as the modalities recorded in each corpus (e.g. text, speech and video). We further discuss corpora constructed from both human-human and human-machine interactions, corpora constructed using natural versus unnatural or constrained settings, and corpora constructed using works of fiction. In Section 4, we present our survey over dialogue corpora using the categories from Sections 2 & 3. In particular, we categorize the corpora based on whether dialogues are: between humans or between a human and a machine; whether the dialogues are in written or spoken language; whether they are constrained or spontaneous; and whether they are scripted fictional works. We discuss each corpus in turn while emphasizing how the dialogues were generated and collected, the topic of the dialogues, and the size of the entire corpus. In Section 5, we discuss issues related to: corpus size, transfer learning between corpora, incorporation of external knowledge into the dialogue system, data-driven learning for contextualization and personalization, and automatic evaluation metrics. We conclude the survey in Section 6.

2. Characteristics of Data-Driven Dialogue Systems

This section offers a broad characterization of data-driven dialogue systems, which structures our presentation of the datasets.

2.1 An Overview of Dialogue Systems

The standard architecture for dialogue systems, shown in Figure 1, incorporates Automatic Speech Recognition, Natural Language Understanding, Dialogue State Tracking, Dialogue Response Action Selection, Natural Language Generation, and Speech Synthesis. In the case of text-based (written) dialogues, Automatic Speech Recognition and Speech Synthesis can be left out and throughout this work we generally do not discuss them and focus on the core components. We can refer to systems including Speech Synthesis and Automatic Speech Recognition as Spoken Dialogue Systems. In some dialogue systems literature, the Dialogue State Tracking and Dialogue Response Action Selection components comprise the Dialogue Manager (Young, 2000). Here, we discuss each component separately and briefly discuss related work to each. We also discuss end-to-end dialogue systems, a relatively new paradigm which ignores the division into components and treats all four non-spoken dialogue system components as a single learned system (Ritter et al., 2011; Vinyals and Le, 2015; Lowe et al., 2015a; Sordoni et al., 2015; Shang et al., 2015; Li et al., 2015; Serban et al., 2016; Serban et al., 2017d,a; Dodge et al., 2015; Williams and Zweig, 2016; Weston, 2016).

In this work, we focus on corpus-based data-driven dialogue systems. That is, systems that use machine learning algorithms to learn the functionality of the previously described components from dialogue corpora constructed from human conversational data. We define human conversational data as dialogue corpora collected through human-human or human-machine interaction. These system components have variables or parameters that are optimized based on statistics observed in dialogue corpora. In particular, we focus on systems where the majority of variables and parameters are optimized. Such corpus-based data-driven systems should be contrasted to systems where each component is hand-crafted by engineers — for example, components defined by an a priori fixed set of deterministic rules (e.g. Weizenbaum (1966); McGlashan et al. (1992)). These systems should also be contrasted with systems learning online, such as when the free variables and parameters are optimized directly based on interactions with humans (e.g. Gašić et al. (2011)). Still, it is worth noting that it is possible to combine different types of learning within one system. For example, some parameters may be learned using statistics observed in a corpus, while other parameters may be learned through interactions with humans.

While there are substantial opportunities to improve each of the components in Figure 1 through (corpus-based) data-driven approaches, within this survey we focus primarily on datasets suitable to jointly enhance the components inside the Dialogue System box. It is worth noting that Natural Language Understanding and Natural Language Generation are core problems in Natural Language Processing with applications well beyond dialogue systems.

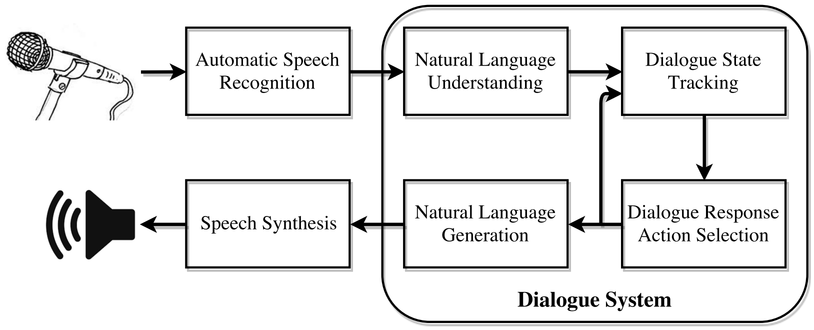

Figure 1: Spoken Dialogue System Diagram. Rectangular boxes represent the different components of the system while arrows represent how the information flows between the components from Automatic Speech Recognition all the way to Speech Synthesis. Throughout this survey we mostly discuss text-based dialogue systems (denoted in the Figure as Dialogue Systems) which omit the Automatic Speech Recognition and Speech Synthesis.

2.2 Tasks and objectives

Dialogue systems have been built for a wide range of purposes. A useful distinction can be made between goal-driven dialogue systems, such as technical support services, and non-goal-driven dialogue systems, such as bots aimed for general chatting (Wallace, 2009; Serban et al., 2017b; A. Ram, 2017; I. Papaioannou, 2017; Serban et al., 2017c). Although both types of systems do in fact have objectives, typically the goal-driven dialogue systems have a well-defined measure of performance that is explicitly related to task completion. While both goal-driven and non-goal-driven components may have different levels of abstraction, throughout this text, we generally refer to the surface-form output of a dialogue system as the dialogue system or natural language response (i.e. the Natural Language Generation output) and a concept-level response (e.g. the output of a Dialogue Manager or Response Action Selection Mechanism) as a response action.

Non-goal-driven Dialogue Systems. Research on non-goal-driven dialogue systems goes back to the mid-60s. It began, perhaps, with Weizenbaum's famous program *ELIZA*, a system based only on simple text parsing rules that managed to convincingly mimic a Rogerian psychotherapist by persistently rephrasing statements or asking questions (Weizenbaum, 1966). This line of research was continued by Colby (1981), who used simple text parsing rules to construct the dialogue system *PARRY*, which managed to mimic the pathological behaviour of a paranoid patient to the extent that clinicians could not distinguish it from real patients. However, neither of these two systems used data-driven learning approaches. Later work, such as the MegaHal system by Hutchens and Alder (1998), started to apply data-driven methods (Shawar and Atwell, 2007b). Hutchens and Alder (1998) proposed modeling dialogue as a stochastic sequence of discrete symbols (words) using 4th-order Markov chains. Given a user utterance, their system generated a response by following a two-step procedure: first, a sequence of topic keywords, used to create a seed reply, was extracted from the user's utterance; second, starting from the seed reply, two separate Markov chains generated the words preceding and proceeding the seed keywords. This procedure produced many candidate responses, from which the highest entropy response was returned to the user. Under the assumption that the coverage of different topics and general fluency is of primary importance, the 4th order Markov chains were trained on a mixture of data sources ranging from real and fictive dialogues to arbitrary texts. Until very recently, such data-driven dialogue systems were not applied widely in real-world applications (Perez-Marin and Pascual-Nieto, 2011; Shawar and Atwell, 2007b). More recently, in a similar spirit, several neural network architectures trained on large-scale corpora have been developed. These models show promising results for several non-goal-driven dialogue tasks (Ritter et al., 2011; Vinyals and Le, 2015; Lowe et al., 2015a; Sordoni et al., 2015; Shang et al., 2015; Li et al., 2015; Serban et al., 2016; Serban et al., 2017d,a; Dodge et al., 2015; Williams and Zweig, 2016; Weston, 2016). However, they require having sufficiently large corpora — in the hundreds of millions or even billions of words — in order to achieve these results.

Goal-driven Dialogue Systems. Initial work on goal-driven dialogue systems was primarily based on deterministic hand-crafted rules coupled with learned speech recognition models. One example is the SUNDIAL project, which was capable of providing timetable information about trains and airplanes, as well as taking airplane reservations (Aust et al., 1995; McGlashan et al., 1992; Simpson and Eraser, 1993). Later, machine learning techniques were used to classify the intention (or need) of the user, as well as to bridge the gap between text and speech (e.g. by taking into account uncertainty related to the outputs of the speech recognition model) (Gorin et al., 1997). Research in this area started to take off during the mid 1990s, when researchers began to formulate dialogue as a sequential decision making problem based on Markov decision processes (Singh et al., 1999; Young et al., 2013; Paek, 2006; Pieraccini et al., 2009). Unlike for non-goal-driven systems, industry played a major role and enabled researchers to have access to (at the time) relatively large dialogue corpora for certain tasks, such as recordings from technical support call centers. Although research in the past decade has continued to push the field towards data-driven approaches, commercial systems are highly domain-specific and heavily based on hand-crafted rules and features (Young et al., 2013). In particular, many of the tasks and datasets available are constrained to narrow domains.

2.3 Learning Dialogue System Components

Most of the Dialogue System components shown in Figure 1 can be learned through so-called discriminative models, which aim to predict labels or annotations relevant to other parts of the dialogue system. Many discriminative models, such as the ones we focus on in this section, fall into the machine learning paradigm of supervised learning. When the labels of interest are discrete (most commonly) the models are called *classification* models. When the labels of interest are continuous, the models are called *regression* models. One popular approach for tackling the discriminative task is to learn a probabilistic model of the labels conditioned on the available information, $P(Y|X)$, where Y is the label of interest (e.g. a discrete variable representing the user intent), and X is the available information (e.g. utterances in the conversation). Another popular approach is to use maximum margin classifiers such as support vector machines (Cristianini and Shawe-Taylor, 2000) as opposed to probabilistic models.

Discriminative models have allowed goal-driven dialogue systems to make significant progress (Williams et al., 2013). With proper annotations, discriminative models can be evaluated automatically and accurately. Furthermore, once trained on a given dataset, these models may be plugged into a fully-deployed dialogue system (e.g. a classification model for user intents may be used as input to Dialogue State Tracking). Although it is beyond the scope of this paper to provide a survey over such system components, we now give a brief example of each component. This will motivate and facilitate the dataset analysis in Section 4.

2.3.1 NATURAL LANGUAGE UNDERSTANDING

A Natural Language Understanding model, placed in the context of a dialogue system, is typically designed to interpret and classify the semantic meaning or intent of the interlocutor. Several works investigate different statistical approaches for learning Natural Language Understanding models. These involve semantic frame parsing Wang et al. (2005), intent classification (Tur and Deng, 2011), slot-filling (Mesnil et al., 2013; Liu and Lane, 2017), and semantic interpretation (Miller et al., 1996).

Discriminative models, as aforementioned, are often used in natural language understanding for user intent classification. This model is trained to predict the intent of a user conditioned on the utterances of that user. In this case, the intent is called the *label* (or *target* or *output*), and the conditioned utterances are called the *conditioning variables* (or *inputs*). Training this model requires examples of pairs of user utterances and intentions. One way to obtain these example pairs would be to first record written dialogues between humans carrying out a task, and then to have humans annotate each utterance with its intention label. Depending on the complexity of the domain, this may require training the human annotators to reach a certain level of agreement between annotators.

Often, Natural Language Understanding systems are more expansive than this. Along with an intent, these systems can also provide so-called slot information (Mesnil et al., 2013; Liu and Lane, 2017). This is added information which narrows the scope of the intent. For example, the system may classify a “Request Arrival Times Intent” with a slot of “Flight Number = AZ1234”.

2.3.2 DIALOGUE MANAGEMENT

As previously mentioned, dialogue management encompasses both Dialogue State Tracking and Dialogue Response Generation. Often, the role of a Dialogue Manager component is to take the current dialogue state (e.g. output from Natural Language Understanding components) and take an action – typically at the concept level – which can be transformed into a natural language response. See Churcher et al. (1997); Young (2000); Lee et al. (2010) for overviews of data-driven dialogue management components using various methods. Recent advances in dialogue management have involved using reinforcement learning methods to learn a dialogue management policy (Singh et al., 2002; Scheffler and Young, 2002; Pietquin et al., 2011; Rieser and Lemon, 2011; Peng et al., 2017; Fazel-Zarandi et al., 2017). An example of such a learned dialogue management policy, as seen in Fazel-Zarandi et al. (2017), takes the intent and slot information outputted by a Natural Language Understanding system, along with the history of this information and confidence score, and outputs one of several actions: confirming what the interlocutor said, eliciting more information from the interlocutor, or selecting a response according to a pre-defined response generation policy. This concept-level action is then converted to an utterance by the Natural Language Generation component.

Dialogue State Tracking. The Dialogue State Tracking component of a dialogue system might similarly be implemented as a classification model (Williams et al., 2013). At any given point in the dialogue, such a model will take as input all the user utterances and user intention labels estimated by a Natural Language Understanding model so far, and outputs a distribution over possible dialogue states. One common way to represent dialogue states are through slot-value pairs. For example, a dialogue system providing timetable information for trains might have three different slots: *departure city, arrival city*, and *departure time*. Each slot may take one of several discrete values (e.g. *departure city* could take values from a list of city names). The task of Dialogue State Tracking is then to output a distribution over every possible combination of slot-value pairs. This distribution — or alternatively, the K dialogue states with the highest probability — may then be used by other parts of the dialogue system. The Dialogue State Tracking model can be trained on examples of dialogue utterances and dialogue states labeled by humans.

Dialogue Response Action Selection. Given the dialogue state distribution provided by the Dialogue State Tracking system, the Dialogue Response Action Selection component must select an appropriate system response action (sometimes referred to as a dialogue act). This component may also be implemented as a classification model that maps dialogue states to a probability over a discrete set of response actions. For example, in a dialogue system providing timetable information for trains, the set of response actions might include *providing information* (e.g. providing the departure time of the next train with a specific departure and arrival city) and *clarification questions* (e.g. asking the user to re-state their departure city). The model may be trained on example pairs of dialogue states and response actions.

2.3.3 NATURAL LANGUAGE GENERATOR

Given a dialogue system selected response action (e.g. a response action providing the departure time of a train), the Natural Language Generator must output the natural language utterance of the system. In the case of commercial goal-driven dialogue systems, this is often implemented using hand-crafted rules. Another option is to learn a discriminative model to select a natural language response. In this case, the output space may be defined as a set of so-called *surface form* sentences (e.g. *"The requested train leaves city X at time Y"*, where X and Y are placeholder values). Given the system response action, the classification model must choose an appropriate surface form. Afterwards, the chosen surface form will have the placeholder values substituted with appropriate items (e.g. X will be replaced by the appropriate city name through a database look up). Several works examine such data-driven approaches. The authors of (Oh and Rudnicky, 2000) use statistical probabilities gathered from corpora to generate a conditional language generation process. Similarly Lemon (2008), reformulates this mapping as a reinforcement learning problem and trains linear policies to generate natural language based on semantic frames. Wen et al. (2015b) use recurrent neural networks to learn a mapping from a dialogue act and semantic frame to a natural language response. Many other works, similarly propose various methods for conditional natural language generation based on response actions, dialogue acts, and semantic frames. As with other classification models, such models may be trained on example pairs of system response actions and surface forms.

2.3.4 END-TO-END DIALOGUE SYSTEMS

Not all dialogue systems conform to the architecture shown in Figure 1. Various works have examined learning "end-to-end" systems which combine various sub-components together. While some works investigate combining various combinations of sub-components together — such as Natural Language Understanding and Dialogue Management (Yang et al., 2017) — we take a broader view where we define end-to-end systems as encompassing all of the non-spoken dialogue components (i.e., Natural Language Understanding, Dialogue State Tracking, Dialogue Response Action Selection, and Natural Language Generation).

In particular, so-called end-to-end dialogue system architectures based on neural networks have shown promising results on several dialogue tasks (Ritter et al., 2011; Vinyals and Le, 2015; Lowe et al., 2015a; Sordoni et al., 2015; Shang et al., 2015; Li et al., 2015; Serban et al., 2016; Serban et al., 2017d,a; Dodge et al., 2015). In their purest form, these models take as input a dialogue in text form and output a natural language response (or a distribution over responses). We call these systems end-to-end dialogue systems because they possess two important properties. First, they do not contain or require learning any sub-components (such as Natural Language Understanding or Dialogue State Tracking). Consequently, there is no need to collect intermediate labels (e.g. user intention or dialogue state labels). Second, all model parameters are optimized w.r.t. a single objective function. Often the objective function chosen is maximum log-likelihood (e.g. cross-entropy) on a fixed corpus of dialogues. Although in earlier work these models depended only on the dialogue context, they may be extended to also depend on outputs from other components (e.g. outputs from the speech recognition component), and on external knowledge (e.g. external databases that can be queried by the system).

End-to-end dialogue systems can be divided into two categories: those that select deterministically from a fixed set of possible responses, and those that attempt to stochastically generate responses by keeping a posterior distribution over possible utterances. Systems in the first category map the dialogue history, state tracking outputs and external knowledge to a response action:

$$f_\theta \ : \{\text{dialogue history},\ \text{state tracking outputs}, \text{external knowledge}, t\} \rightarrow \ \text{action } a_t, \qquad (1)$$

where a_t is the dialogue system response action at time t, and θ is the set of parameters that defines f. While the goal of end-to-end systems is to output the response directly (for example, taking a word output as an action), in current systems, the response action may also refer to a selection from a pre-defined set of linguistic responses — perhaps corresponding to different dialogue acts. Information retrieval and ranking-based systems — systems that search through a database of dialogues and pick responses with the most similar context, such as the model proposed by Banchs and Li (2012) — belong to this category. In this case, the mapping function f_θ projects the dialogue history into a Euclidean space (e.g. using TF-IDF bag-of-words representations). The response is then found by projecting all potential responses into the same Euclidean space, and the response closest to the desirable response region is selected. The neural network proposed by Lowe et al. (2015a) also belongs to this category. In this case, the dialogue history is projected into a Euclidean space using a recurrent neural network encoding the dialogue word-by-word. Similarly, a set of candidate responses are mapped into the same Euclidean space using another recurrent neural network encoding the response word-by-word. Finally, a relevance score is computed between the dialogue context and each candidate response, and the response with the highest score is returned. Hybrid or combined models, such as the model built on both a phrase-based statistical machine translation system and a recurrent neural network proposed by Sordoni et al. (2015), also belong to this category. In this case, a response is generated by deterministically creating a fixed number of answers using the machine translation system and then picking the response according to the score given by a neural network. Although both of its sub-components are based on probabilistic models, the final model does not construct a probability distribution over all possible responses.[1]

1. Although the model does not require intermediate labels, it consists of sub-components whose parameters are trained with different objective functions. Therefore, strictly speaking, this is not an end-to-end model.

In contrast to a deterministic system, a stochastic system explicitly computes a full posterior probability distribution over possible system response actions at every turn:

$$P_\theta(\text{action } a_t \mid \text{dialogue history, state tracking outputs, external knowledge}, t). \quad (2)$$

Systems based on generative recurrent neural networks typically belong to this category (Vinyals and Le, 2015; Serban et al., 2016). By breaking down Eq. (2) into a product of probabilities over words, responses can be generated by sampling word-by-word from their probability distribution. These systems are also able to generate entirely novel responses by sampling word-by-word (though, some such models require modification to elicit diversity in responses (Li et al., 2015)). Highly probable responses, i.e. the response with the highest probability, can further be generated by using a method known as beam-search (Graves, 2012). These systems project each word into a Euclidean space (known as a word embedding) (Bengio et al., 2003); they also project the dialogue history and external knowledge into a Euclidean space (Wen et al., 2015a; Lowe et al., 2015b). Similarly, the system proposed by Ritter et al. (2011) belongs to this category. Their model uses a statistical machine translation model to map a dialogue history to its response. When trained solely on text, these generative models can be viewed as unsupervised learning models, because they aim to reproduce the training data distribution. In other words, such models learn to assign a probability to every possible conversation, and since they generate responses word-by-word, they must learn to simulate the behaviour of the agents in the training corpus.

Early reinforcement learning dialogue systems with stochastic policies also belong to this category (e.g. the NJFun system of Singh et al. (2002)). In contrast to the neural network and statistical machine translation systems, these reinforcement learning systems typically have very small sets of possible hand-crafted system states (e.g. hand-crafted features describing the dialogue state). The action space is also limited to a small set of pre-defined responses. This makes it possible to apply established reinforcement learning algorithms to train them either online or offline, however it also severely limits their application area. As Singh et al. (Singh et al., 2002, p.5) remark: *"We view the design of an appropriate state space as application-dependent, and a task for a skilled system designer."*

3. Dialogue Interaction Types & Aspects

This section provides a high-level discussion of different types of dialogue interactions and their salient aspects. The categorization of dialogues is useful for understanding the utility of various datasets for particular applications, as well as for grouping these datasets together to demonstrate available corpora in a given area. Here, we discuss the important characteristics and distinctions within dialogue corpora: whether a corpus is written, spoken, or multi-modal; whether a corpus features human-human interactions or human-machine interactions; whether the corpus features constrained or unconstrained dialogues; whether the corpus includes topic oriented or goal driven dialogues; whether the dialogues are fictional or scripted; corpus size.

3.1 Written, Spoken & Multi-modal Corpora

An important distinction between dialogue corpora is whether participants (interlocutors) interact through written language, spoken language, or in a multi-modal setting (e.g. using both speech and visual modalities). Written and spoken language differ substantially w.r.t. their linguistic properties.

Spoken language tends to be less formal, containing lower information content and many more pronouns than written language (Carter and McCarthy, 2006; Biber and Finegan, 2001, 1986). In particular, the differences are magnified when written language is compared to spoken face-to-face conversations, which are multi-modal and highly socially situated. As Biber and Finegan (1986) observed, pronouns, questions, and contradictions, as well as that-clauses and if-clauses, appear with a high frequency in face-to-face conversations. Forchini (2012) summarized these differences: *"... studies show that face-to-face conversation is interpersonal, situation-dependent has no narrative concern or as Biber and Finegan (1986) put it, is a highly interactive, situated and immediate text type..."* Due to these differences between spoken and written language, we emphasize the distinction between dialogue corpora in written and spoken language in the following sections.

Similarly, dialogues involving visual and other modalities differ from dialogues without these modalities (Card et al., 1983; Goodwin, 1981). When a visual modality is available —- for example, when two human interlocutors converse face-to-face — body language and eye gaze have a significant impact on what is said and how it is said (Gibson and Pick, 1963; Lord and Haith, 1974; Cooper, 1974; Chartrand and Bargh, 1999; de Kok et al., 2013). Aside from the visual modality, dialogue systems may also incorporate other situational modalities, including aspects of virtual environments (Rickel and Johnson, 1999; Traum and Rickel, 2002) and user profiles (Li et al., 2016).

3.2 Human-Human Vs. Human-Machine Corpora

Another salient distinction between dialogue datasets resides in the types of interlocutors — notably, whether it involves interactions between two humans, or between a human and a computer.[2] The distinction is important because current artificial dialogue systems are significantly constrained. These systems do not produce nearly the same distribution of possible responses as humans do under equivalent circumstances. As stated by Williams and Young (2007):

> (Human-human conversation) does not contain the same distribution of understanding errors, and human-human turn-taking is much richer than human-machine dialog. As a result, human-machine dialogue exhibits very different traits than human-human dialogue (Doran et al., 2001; Moore and Browning, 1992).

The expectation a human interlocutor begins with, and the interface through which they interact, also affect the nature of the conversation (Jonsson and Dahlback, 1988).

For goal-driven settings, Williams and Young (2007) have argued against building data-driven dialogue systems using human-human dialogues, as it contains a different distribution of understanding errors. This line of reasoning seems particularly applicable to spoken dialogue systems, where speech recognition errors can have a critical impact on performance and therefore must be taken into account when learning the dialogue model. The argument is also relevant to goal-driven dialogue systems, where an effective dialogue model can often be learned using reinforcement learning techniques. Williams and Young (2007) also argue against learning from corpora generated between humans and existing dialogue systems, as the trained dialogue system would simply learn to approximate the policy of the spoken dialogue system.

Thus, if one's goal is to develop a dialogue system that can interact with real users, the most effective strategy may be learning online through interaction with the users. For example, there

2. Machine-machine dialogue corpora are not of interest to us, because they typically differ significantly from natural human language. Furthermore, user simulation models are outside the scope of this survey.

exists useful human-machine corpora where the interacting machine uses a stochastic policy that can generate sufficient coverage of the task (e.g. enough *good* and enough *bad* dialogue examples) to allow an effective dialogue model to be learned. In this case, the goal is to learn a policy that is eventually better than the original stochastic policy used to generate the corpus through a process known as bootstrapping. Furthermore, there may be other reasons to prefer human-machine over human-human corpora, for example if researchers desire to study the behavior of a particular dialogue system.

Another possible alternative is the case of *Wizard-of-Oz* experiments (Bohus and Rudnicky, 2008; Petrik, 2004). In these dialogue collection methods, a human thinks (s)he is speaking to a machine, but a human operator is in fact controlling the dialogue system. This enables the generation of datasets that are closer in nature to the dialogues humans may wish to achieve in a given setting. For such experiments, it may also be beneficial to influence either or both the human wizard and user to collect a diverse dataset. For example, in a negotiation task Konovalov et al. (2016) propose to 1) influence the human wizard by asking them to change their negotiation strategy, and 2) have the human user rephrase their utterances for low-frequency intentions. However, *Wizard-of-Oz* experiments are typically expensive and time-consuming to carry out.

3.3 Spontaneous Vs. Constrained Corpora

The way in which a dialogue corpus is generated and collected can have a significant influence on the trained data-driven dialogue system. Many dialogue corpora contain dialogues in which the topics of conversation are either casual or not pre-specified in any way. These corpora can be referred to as *Spontaneous (Unconstrained) Corpora*, as we believe they most closely mimic spontaneous and unplanned spoken interactions between humans. However, in some corpora, conversations focus on a particular topic or intend to solve a specific task. In such situations, the task or topic is pre-specified and participants are discouraged from deviating from the topic. We refer to these as *Constrained Dialogue Corpora*.

Spontaneous Corpora bear a close resemblance to natural dialogues — that is, they are close to the generally unplanned nature of most spoken interactions between humans. In the latter case — that of constrained dialogues — some experimental conditions in which dialogues were collected can result in unnatural behaviours that do not correlate well to the true typical dialogue patterns of human-human interaction in day-to-day settings. Due to ethical considerations and resource constraints, researchers may be forced to inform the human interlocutors that they are being recorded or to setup artificial experiments in which they hire humans and instruct them to carry out a particular task by interacting with a dialogue system. In these cases, there is no guarantee that the interactions in the corpus will reflect natural human interactions, since the hired humans may behave differently from the population. One factor that may cause behavioural differences is the fact that the hired humans may not share the same intentions and motivations as the population (Ai et al., 2007; Young et al., 2013). The unnaturalness may be further exacerbated by the hiring process, as well as the platform through which they interact. Such factors are becoming more prevalent as researchers increasingly rely on crowdsourcing platforms, such as Amazon Mechanical Turk, to collect and evaluate dialogue data (Jurcıcek et al., 2011).

While these factors concerning constrained corpora are important to consider, there are many use cases where such corpora are beneficial or necessary. Certain researchers may prefer laboratory-like conditions to study certain variables of interest in a conversation. Further, constrained corpora in the form of debate settings, topical discussions, etc. are useful to study in and of themselves. In Sections 4.2 and 4.3, we separate our discussion of dialogue datasets based on whether the corpora are constrained or spontaneous.

3.4 Topic-oriented & Goal-driven Datasets

Many human-human datasets may be described as containing *casual* or *unrestricted* topics, while human-machine datasets often focus on specific, narrow topics. It is useful to keep this distinction between *restricted* and *unrestricted* topics in mind, since goal-driven dialogue systems — which often have a well-defined measure of performance related to task completion — are usually developed in the former setting. When the corpus domain is *restricted* and a completion metric is available, it may be useful to incorporate this explicitly into the learning procedure. In contrast, when building non-goal-driven dialogue systems based on a corpus with *unrestricted* topics, it may not be possible to explicitly incorporate any topic information or completion metric into the learning procedure.

In some cases, the line between *restricted* and *unrestricted* topics blurs. For example, in the case of conversations between players of an online game (Afantenos et al., 2012), the outcome of the game is determined by how participants play in the game environment, not by their conversation. In this case, some conversations may have a direct impact on a player's performance in the game. Other conversations may be related to the game but irrelevant to the goal (e.g. commentary on past events). Finally, some conversations may be completely unrelated to the game.

3.5 Scripted Corpora or Corpora from Fiction

It is also possible to use artificial dialogue corpora for data-driven learning. This includes corpora based on works of fiction such as novels, movie manuscripts and audio subtitles. However, unlike transcribed human-human conversations, novels, movie manuscripts, and audio subtitles depend upon events outside the current conversation, which are not observed. This makes data-driven learning more difficult because the dialogue system has to account for unknown factors. The same problem is also observed in certain other media, such as microblogging websites (e.g. Twitter and Weibo), where conversations also may depend on external unobserved events.

Nevertheless, recent studies have found that spoken language in movies resembles spontaneous human spoken language (Forchini, 2009). Although movie dialogues are explicitly written to be spoken and contain certain artificial elements, many of the linguistic and paralinguistic features contained within the dialogues are similar to natural spoken language, including dialogue acts such as turn-taking and reciprocity (e.g. returning a greeting when greeted). The artificial differences that exist may even be helpful for data-driven dialogue learning since movie dialogues are more compact, follow a steady rhythm, and contain less garbling and repetition, all while still presenting a clear event or message to the viewer (Dose, 2013; Forchini, 2009, 2012). Unlike dialogues extracted from *Wizard-of-Oz* human experiments, movie dialogues span many different topics and occur in many different environments (Webb, 2010). They contain different actors with varying intentions and relationships to one another, which could potentially allow a data-driven dialogue system to learn to personalize itself to each user by identifying different interaction patterns (Li et al., 2016).

3.6 Corpus Size

As in other machine learning applications such as machine translation (Al-Onaizan et al., 2000; Gülçehre et al., 2015) and speech recognition (Deng and Li, 2013; Bengio et al., 2014), the size of the dialogue corpus is important for building an effective data-driven dialogue system (Lowe et al., 2015a; Serban et al., 2016).

There are two primary perspectives on the importance of dataset size for building data-driven dialogue systems. The first perspective comes from the machine learning literature: larger datasets place constraints on the dialogue model trained from that data. Datasets with few examples typically require strong structural priors placed on the model, such as using a modular system, whereas large datasets can be used to train end-to-end dialogue systems with less *a priori* structure. The second comes from a statistical natural language processing perspective: since the statistical complexity of a corpus grows with the linguistic diversity and number of topics, the number of examples required by a machine learning algorithm to model the patterns in it will also grow with the linguistic diversity and number of topics. Consider two small datasets with the same number of dialogues in the domain of bus schedule information: in one dataset the conversations between the users and operator is natural, and the operator can improvise and chitchat; in the other dataset, the operator reads from a script to provide the bus information. Despite having the same size, the second dataset will have less linguistic diversity and not include chitchat topics. Therefore, it will be easier to train a data-driven dialogue system mimicking the behaviour of the operator in the second dataset, however it will also exhibit a highly pedantic style and not be able to chitchat. In addition to this, to have an effective discussion between any two agents, their common knowledge must be represented and understood by both parties. The process of establishing this common knowledge, also known as *grounding*, is especially critical to repair misunderstandings between humans and dialogue systems (Cahn and Brennan, 1999). Since the number of misunderstandings can grow with the lexical diversity and number of topics (e.g. misunderstanding the paraphrase of an existing word, or misunderstanding a rarely seen keyword), the number of examples required to repair these grow with linguistic diversity and topics. In particular, the effect of linguistic diversity has been observed in practice: Vinyals and Le (2015) train a simple encoder-decoder neural network on a proprietary dataset of technical support dialogues. Although it has a similar size and purpose as the Ubuntu Dialogue Corpus (Lowe et al., 2015a), the qualitative examples shown by Vinyals and Le (2015) are significantly superior to those obtained by more complex models on the Ubuntu Corpus (Serban et al., 2017a). This result may likely be explained in part due to the fact that technical support operators often follow a comprehensive script for solving problems. As such, the script would reduce the linguistic diversity of their responses.

Furthermore, since the majority of human-human dialogues are multi-modal and highly ambiguous in nature (Chartrand and Bargh, 1999; de Kok et al., 2013), the size of the corpus may compensate for some of the ambiguities and missing modalities. That is, humans express themselves in conversation through intonation, emotional undertones, contextual cues, body language, and other aspects which may not be conveyed fully in dialogue corpora. If the corpus is sufficiently large, then the resolved ambiguities and missing modalities may, for example, be approximated using latent stochastic variables (Serban et al., 2017d). Thus, we include corpus size as a dimension of analysis. We also discuss the benefits and drawbacks of several popular large-scale datasets in Section 5.1.

4. Available Dialogue Datasets

There is a vast amount of data available documenting human communication. Much of this data could be used — perhaps after some pre-processing — to train a dialogue system. However, covering all such sources of data would be infeasible. Thus, we restrict the scope of this survey to datasets that have already been used to study dialogue or build dialogue systems or which could be leveraged in the near future to build more sophisticated data-driven dialogue models (due to properties of dataset size or annotations useful for different data-driven dialogue components). We restrict the selection further to contain only corpora generated from spoken or written English, and to corpora which, to the best of our knowledge, either are publicly available or will be made available in the near future. We first give a brief overview of each of the considered corpora. Later, we highlight promising examples and explain how they could be used to further dialogue research.[3]

The dialogue datasets analyzed are listed in Tables 1–5. Column features indicate properties of the datasets, including the number of dialogues, average dialogue length, number of words, whether the interactions are between humans or with an automated system, and whether the dialogues are written or spoken. Below, we discuss qualitative features of the datasets, while statistics can be found in the aforementioned table. We generally divide the corpora according to the main characteristics mentioned previously: Human-Machine Corpora (including spoken and written systems), Human-Human Spontaneous (Unconstrained) Spoken Corpora, Human-Human Constrained Spoken Corpora, Human-Human Scripted (Mostly Fictional) Spoken Corpora, Human-Human Spontaneous (Unconstrained) Written Corpora, and Human-Human Constrained Written Corpora.

4.1 Human-Machine Corpora

As discussed in Subsection 3.2, an important distinction between dialogue datasets is whether they consist of dialogues between two humans or between a human and a machine. Thus, we begin by outlining some of the existing human-machine corpora in several categories based on the types of systems the humans interact with: Restaurant and Travel Information, Open-Domain Knowledge Retrieval, and Other Specialized systems. Note, we also include human-human corpora here where one human plays the role of the machine in a Wizard-of-Oz fashion.

4.1.1 RESTAURANT AND TRAVEL INFORMATION

One common theme in human-machine language datasets is interaction with systems that provide restaurant or travel information. Here we briefly describe some human-machine dialogue datasets in this domain.

One of the early and most influential sources of such data is the **Let's Go!** dataset (Raux et al., 2005), which was collected from people calling a bus scheduling service at off-peak times.[4] The dataset provides over 170,000 conversations recorded between an automated bus information system and callers requesting bus schedule information.

One of the more recent sources of such data has come from the datasets for structured dialogue prediction released in conjunction with the **Dialog State Tracking Challenge** (DSTC) (Williams et al., 2013). As the name implies, these datasets are used to learn a strategy for Dialogue State

3. We form a live list of the corpora discussed in this work, along with links to downloads, at: `http://breakend.github.io/DialogDatasets`. Pull requests can be made to the Github repository (`https://github.com/Breakend/DialogDatasets`) hosting the website for continuing updates to the list of corpora.

4. See `https://dialrc.github.io/LetsGoDataset/`.

Tracking (sometimes called "belief tracking"), which involves estimating the intentions of a user throughout a dialog. State tracking is useful as it can increase the robustness of speech recognition systems, and can provide an implementable framework for real-world dialogue systems. Particularly in the context of goal-oriented dialogue systems (such as those providing travel and restaurant information), state tracking may be necessary for creating coherent conversational interfaces. That is, to form a coherent dialogue, previous contexts must be accounted for – either explicitly or in an end-to-end manner. As such, the first three datasets in the DSTC — referred to as **DSTC1, DSTC2**, and **DSTC3** respectively — are medium-sized spoken datasets obtained from human-machine interactions with restaurant and travel information systems. All datasets provide labels specifying the current goal and desired action of the system. **DSTC1** (Williams et al., 2013) is an annotated subset of the conversations from the **Let's Go!** dataset (Parent and Eskenazi, 2010) discussed earlier, involving conversations between callers and an automated bus information system. **DSTC2** introduces changing user goals in a restaurant booking system, while trying to provide a desired reservation (Henderson et al., 2014b). **DSTC3** introduces a small amount of labeled data in the domain of tourist information. It is intended to be used in conjunction with the **DSTC2** dataset as a domain adaptation problem (Henderson et al., 2014a).

The **Carnegie Mellon Communicator Corpus** (Bennett and Rudnicky, 2002) also contains human-machine interactions with a travel booking system. It is a medium-sized dataset of interactions with a system providing up-to-the-minute flight information, hotel information, and car rentals. Conversations with the system were transcribed, along with user's comments after the interaction.

The **ATIS (Air Travel Information System) Pilot Corpus** (Hemphill et al., 1990) is one of the first human-machine corpora. It consists of interactions, lasting about 40 minutes each, between human participants and a travel-type booking system, secretly operated by humans. This dataset contains 1041 utterances.

In the **Maluuba Frames Corpus** (El Asri et al., 2017), one user plays the role of a conversational agent in a Wizard-of-Oz fashion, while the other user is tasked with finding available travel or vacation accommodations according to a pre-specified task. The Wizard is provided with a knowledge database which records its actions. Semantic frames are annotated in addition to actions which the Wizard performed on the database to accompany a line of dialogue. In this way, the Frames corpus aims to track decision-making processes in travel- and hotel-booking through natural dialog.

4.1.2 OPEN-DOMAIN KNOWLEDGE RETRIEVAL

Knowledge retrieval and Question & Answer (QA) corpora are a broad distinction of corpora that we will not extensively review here. Instead, we include only those QA corpora which explicitly record interactions of humans with existing systems. The **Ritel corpus** (Rosset and Petel, 2006) is a small dataset of 528 dialogues with the Wizard-of-Oz Ritel platform. The project's purpose was to integrate spoken language dialogue systems with open-domain information retrieval systems, with the end goal of allowing humans to ask general questions and iteratively refine their search. The questions in the corpus mostly revolve around politics and the economy (e.g. "Who is currently presiding the Senate?"), along with some conversations about arts- and science-related topics.

Other similar open-domain corpora in this area include WikiQA Yang et al. (2015) and MS MARCO Nguyen et al. (2016), which compile responses from automated Bing searches and human annotators. These do not record dialogues, but rather gather possible responses to queries. We will only mention them briefly as examples of other Open-Domain corpora in the field.

Name	Type	Topics	Avg. # of turns	Total # of dialogues	Total # of words	Description
Let's Go! (Raux et al., 2005)	Spoken	Bus schedules	–	171,128	–	Bus ride information system
DSTC1 (Williams et al., 2013)	Spoken	Bus schedules	13.56	15,000	3.7M	Bus ride information system
DSTC2 (Henderson et al., 2014b)	Spoken	Restaurants	7.88	3,000	432K	Restaurant booking system
DSTC3 (Henderson et al., 2014a)	Spoken	Tourist information	8.27	2,265	403K	Information for tourists
CMU Communicator Corpus (Bennett and Rudnicky, 2002)	Spoken	Travel	11.67	15,481	2M*	Travel planning and booking system
ATIS Pilot Corpus† (Hemphill et al., 1990)	Spoken	Travel	25.4	41	11.4K*	Travel planning and booking system
Ritel Corpus† (Rosset and Petel, 2006)	Spoken	Unrestricted/ Diverse Topics	9.3*	582	60k	An annotated open-domain question answering spoken dialogue system
DIALOG Mathematical Proofs (Wolska et al., 2004)	Spoken	Mathematics	12	66	8.7K*	Humans interact with computer system to do mathematical theorem proving
MATCH Corpus† (Georgila et al., 2010)	Spoken	Appointment Scheduling	14.0	447	69K*	A system for scheduling appointments. Includes dialogue act annotations
Maluuba Frames† (El Asri et al., 2017)	Chat, QA & Recommendation	Travel & Vacation Booking	15	1369	–	For goal-driven dialogue systems. Semantic frames labeled and actions taken on a knowledge-base annotated.

Table 1: Human-machine dialogue datasets. Starred (*) numbers are approximated based on the average number of words per utterance. Datasets marked with (†) indicate Wizard-of-Oz dialogues, where the machine is secretly operated by a human.

4.1.3 OTHER

The **DIALOG mathematical proof dataset** (Wolska et al., 2004) is a Wizard-of-Oz dataset involving an automated tutoring system that attempts to advise students on proving mathematical theorems. This is done using a hinting algorithm that provides clues when students come up with an incorrect answer. At only 66 dialogues, the dataset is very small, and consists of a conglomeration of text-based interactions with the system, as well as think-aloud audio and video footage recorded by the users as they interacted with the system. The latter was transcribed and annotated with simple speech acts such as "signaling emotions" or "self-addressing".

The **MATCH corpus** (Georgila et al., 2010) is a small corpus of 447 dialogues based on a Wizard-of-Oz experiment, which contains conversations from 50 young and old adults interacting with spoken dialogue systems. These conversations were annotated semi-automatically with dialogue acts and "Information State Update" (ISU) representations of dialogue context. The corpus also contains information about the users' cognitive abilities, with the motivation of modeling how the elderly interact with dialogue systems.

4.2 Human-Human Spoken Corpora

Naturally, there is much more data available for conversations between humans than conversations between humans and machines. Thus, we break down this category further, into spoken dialogues (this section) and written dialogues (Section 4.3). The distinction between spoken and written dialogues is important, since the distribution of utterances changes dramatically according to the nature of the interaction. As discussed in Subsection 3.1, spoken dialogues can potentially be less focused as the user speaks in train-of-thought manner. They also tend to use shorter words and phrases. Conversely, in written communication, users have the ability to reflect on what they are writing before they send a message and thus are more precise. Written dialogues can also contain spelling errors or abbreviations; such artifacts are usually not present in datasets transcribed from spoken dialogues.

4.2.1 SPONTANEOUS SPOKEN CORPORA

We first introduce datasets in which the topics of conversation are either casual or not pre-specified in any way. We refer to these corpora as *spontaneous*, as we believe they most closely mimic spontaneous, unplanned spoken interactions between humans.

Perhaps one of the most influential spoken corpora is the **Switchboard dataset** (Godfrey et al., 1992). This dataset consists of approximately 2,500 dialogues from phone calls, along with word-by-word transcriptions, with about 500 different speakers. A computer-driven robot operator system introduced a topic for discussion between two participants, and recorded the resulting conversation. About 70 casual topics were provided, of which about 50 were frequently used. The corpus was originally designed for training and testing various speech processing algorithms; however, it has since been used for a wide variety tasks, including the modeling of dialogue acts such as 'statement', 'question', and 'agreement' (Stolcke et al., 2000).

Another important dataset is the **British National Corpus** (BNC) (Leech, 1992), which contains approximately 10 million words of dialogue. These were collected in a variety of contexts ranging from formal business or government meetings, to radio shows and phone-ins. Although most of the conversations are spoken in nature, some of them are also written. BNC covers a large number of sources, and was designed to represent a wide cross-section of British English from the late twentieth century. The corpus also includes part-of-speech (POS) tagging for every word. The vast array of settings and topics covered by this corpus renders it very useful as a general-purpose spoken dialogue dataset.

Other datasets have been collected for the analysis of spoken English over the telephone. The **CALLHOME American English Speech Corpus** (Canavan et al., 1997) consists of 120 such conversations totalling about 60 hours, mostly between family members or close friends. Similarly, the **CALLFRIEND American English-Non-Southern Dialect Corpus** (Canavan and Zipperlen, 1996) consists of 60 telephone conversations lasting between 5 and 30 minutes each between English speakers in North America without a Southern accent. It is annotated with speaker information such as gender, age, and education. The goal of the project was to support the development of language identification technologies, yet, there are no distinguishing features in either of these corpora in terms of the topics of conversation.

An attempt to capture exclusively teenage spoken language was made in the **Bergen Corpus of London Teenager Language** (COLT) (Haslerud and Stenström, 1995). Conversations were recorded surreptitiously by student 'recruits', with a Sony Walkman and a lapel microphone, in order to obtain a better representation of teenager interactions 'in-the-wild'. This dataset has been used to identify trends in language evolution in teenagers (Stenström et al., 2002).

The **Cambridge and Nottingham Corpus of Discourse in English** (CANCODE) (McCarthy, 1998) is a subset of the Cambridge International Corpus, containing about 5 million words collected from recordings made throughout the islands of Britain and Ireland. It was constructed by Cambridge University Press and the University of Nottingham using dialogue data on general topics between 1995 and 2000. It focuses on interpersonal communication in a range of social contexts, varying from hair salons, to post offices, to restaurants. This has been used, for example, to study language awareness in relation to spoken texts and their cultural contexts (Carter, 1998). In the dataset, the relationships between speakers (e.g. roommates, strangers) are labeled and the interaction types are provided (e.g. professional, intimate).

Other works have attempted to record the physical elements of conversations between humans. To this end, a small corpus entitled **d64 Multimodal Conversational Corpus** (Oertel et al., 2013) was collected, incorporating data from 7 video cameras, and the registration of 3-D head, torso, and arm motion using an Optitrack system. Significant effort was made to make the data collection process as non-intrusive — and thus, natural — as possible. Annotations were made to attempt to quantify overall group excitement and pairwise social distance between participants.

A similar attempt to incorporate computer vision features was made in the **AMI Meeting Corpus** (Renals et al., 2007), where cameras, a VGA data projector capture, whiteboard capture, and digital pen capture, were all used in addition to speech recordings for various meeting scenarios. As with the d64 corpus, the AMI Meeting Corpus is a dataset of multi-participant chats (four-party dialogues) where all members of the party interact with one another. It has been used for analysis of the dynamics of various corporate and academic meeting scenarios, such as addressee detection in a multi-party chat (Akker and Traum, 2009).

In a similar vein, the **Cardiff Conversation Database** (CCDb) (Aubrey et al., 2013) is an audio-visual database containing unscripted natural conversations between pairs of people. The original dataset consisted of 30 five-minute conversations, 7 of which were fully annotated with transcriptions and behavioural annotations such as speaker activity, facial expressions, head motions, and smiles. The content of the conversation is an unconstrained discussion on topics such as movies. While the original dataset featured 2D visual feeds, an updated version with 3D video has also been derived, called the **4D Cardiff Conversation Database** (4D CCDb) (Vandeventer et al., 2015). This version contains 17 one-minute conversations from 4 participants on similarly un-constrained topics.

The **Diachronic Corpus of Present-Day Spoken English** (DCPSE) (Aarts and Wallis, 2006) is a parsed corpus of spoken English made up of two separate datasets. It contains more than 400,000 words from the ICE-GB corpus (collected in the early 1990s) and 400,000 words from the London-Lund Corpus (collected from the late 1960s to the early 1980s). ICE-GB refers to the British component of the International Corpus of English (Greenbaum and Nelson, 1996; Greenbaum, 1996) and contains both spoken and written dialogues from English adults who have completed secondary education. The dataset was selected to provide a representative sample of British English. The London-Lund Corpus (Svartvik, 1990) consists exclusively of spoken British conversations, both dialogues and monologues. It contains a selection of face-to-face, telephone, and public discussion dialogues; the latter refers to dialogues that are heard by an audience that does not participate in the dialogue, including interviews and panel discussions that have been broadcast. The orthographic transcriptions of the datasets are normalized and annotated according to the same criteria; ICE-GB was used as a gold standard for the parsing of DCPSE.

The **Spoken Corpus of the Survey of English Dialects** (Beare and Scott, 1999) consists of 1000 recordings, with about 0.8 million total words, collected from 1948 to 1961 in order to document various existing English dialects. People aged 60 and over were recruited, being most likely to speak the traditional 'uncontaminated' dialects of their area and encouraged to talk about their memories, families, work, and their countryside folklore.

The **Child Language Data Exchange System** (CHILDES) (MacWhinney and Snow, 1985) is a database organized for the study of first and second language acquisition. The database contains 10 million English words and approximately the same number of non-English words. It also contains transcripts, with occasional audio and video recordings of data collected from children and adults learning both first and second languages, although the English transcripts are mostly from children. This corpus could be leveraged in order to build automated teaching assistants.

The expanded **Charlotte Narrative and Conversation Collection** (CNCC), a subset of the first release of the American National Corpus (Reppen and Ide, 2004), contains 95 narratives, conversations and interviews representative of the residents of Mecklenburg County, North Carolina and its surrounding communities. The purpose of the CNCC was to create a corpus of conversation and conversational narration in a 'New South' city at the beginning of the 21st century, that could be used as a resource for linguistic analysis. It was originally released as one of several collections in the New South Voices corpus, which otherwise contained mostly oral histories. Information on speaker age and gender in the CNCC is included in the header for each transcript.

4.2.2 CONSTRAINED SPOKEN CORPORA

Next, we discuss domains in which conversations only occur about a particular topic, or towards solving a specific task. Not only is the topic of the conversation specified beforehand, but participants are discouraged from deviating off-topic. As a result, these corpora are slightly less general than their spontaneous counterparts; however, they may be useful for building goal-oriented dialogue systems. As discussed in Subsection 3.3, this may also make the conversations less natural. We can further subdivide this category into the types of topics they cover: path-finding or planning tasks, persuasion tasks or debates, Q&A or information retrieval tasks, and miscellaneous topics.

Name	Topics	Total # of dialogues	Total # of words	Total length	Description
Switchboard (Godfrey et al., 1992)	Casual Topics	2,400	3M	300hrs*	Telephone conversations on pre-specified topics
British National Corpus (BNC) (Leech, 1992)	Casual Topics	854	10M	1,000hrs*	British dialogues many contexts, from formal business or government meetings to radio shows and phone-ins.
CALLHOME American English Speech (Canavan et al., 1997)	Casual Topics	120	540k*	60hrs	Telephone conversations between family members or close friends.
CALLFRIEND American English Non-Southern Dialect (Canavan and Zipperlen, 1996)	Casual Topics	60	180k*	20hrs	Telephone conversations between Americans with a Southern accent.
The Bergen Corpus of London Teenage Language (Haslerud and Stenström, 1995)	Unrestricted	100	500k	55hrs	Spontaneous teenage talk recorded in 1993. Conversations were recorded secretly.
The Cambridge and Nottingham Corpus of Discourse in English (McCarthy, 1998)	Casual Topics	–	5M	550hrs*	British dialogues from wide variety of informal contexts, such as hair salons, restaurants, etc.
D64 Multimodal Conversation Corpus (Oertel et al., 2013)	Unrestricted	2	70k*	8hrs	Several hours of natural interaction between a group of people
AMI Meeting Corpus (Renals et al., 2007)	Meetings	175	900k*	100hrs	Face-to-face meeting recordings.
Cardiff Conversation Database (CCDb) (Aubrey et al., 2013)	Unrestricted	30	20k*	150min	Audio-visual database with unscripted natural conversations, including visual annotations.
4D Cardiff Conversation Database (4D CCDb) (Vandeventer et al., 2015)	Unrestricted	17	2.5k*	17min	A version of the CCDb with 3D video
The Diachronic Corpus of Present-Day Spoken English (Aarts and Wallis, 2006)	Casual Topics	280	800k	80hrs*	Selection of face-to-face, telephone, and public discussion dialogue from Britain.
The Spoken Corpus of the Survey of English Dialects (Beare and Scott, 1999)	Casual Topics	314	800k	60hrs	Dialogue of people aged 60 or above talking about their memories, families, work and the folklore of the countryside from a century ago.
The Child Language Data Exchange System (MacWhinney and Snow, 1985)	Unrestricted	11K	10M	1,000hrs*	International database organized for the study of first and second language acquisition.
The Charlotte Narrative and Conversation Collection (CNCC) (Reppen and Ide, 2004)	Casual Topics	95	20K	2hrs*	Narratives, conversations and interviews representative of the residents of Mecklenburg County, North Carolina.

Table 2: Human-human spontaneous spoken dialogue datasets. Starred (*) numbers are estimates based on the average rate of English speech from the National Center for Voice and Speech (www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html)

Collaborative Path-Finding or Planning Tasks Several corpora focus on task planning or path-finding through the collaboration of two interlocutors. In these corpora typically one person acts as the decision maker and the other acts as the observer.

A well-known example of such a dataset is the **HCRC Map Task Corpus** (Anderson et al., 1991), that consists of unscripted, task-oriented dialogues that have been digitally recorded and transcribed. The corpus uses the Map Task (Brown et al., 1984), where participants must collaborate verbally to reproduce a route from one of the participant's map to the map of another participant. The corpus is fairly small, but it controls for the familiarity between speakers, eye contact between speakers, matching between landmarks on the participants' maps, opportunities for contrastive stress, and phonological characteristics of landmark names. By adding these controls, the dataset attempts to focus on solely the dialogue and human speech involved in the planning process.

The **Walking Around Corpus** (Brennan et al., 2013) consists of 36 dialogues between people communicating over mobile telephone. The dialogues have two parts: first, a 'stationary partner' is asked to direct a 'mobile partner' to find 18 destinations on a medium-sized university campus. The stationary partner is equipped with a map marked with the target destinations accompanied by photos of the locations, while the mobile partner is given a GPS navigation system and a camera to take photos. In the second part, the participants are asked to interact in-person in order to duplicate the photos taken by the mobile partner. The goal of the dataset is to provide a testbed for natural lexical entrainment, and to be used as a resource for pedestrian navigation applications.

The **TRAINS 93 Dialogues Corpus** (Heeman and Allen, 1995) consists of recordings of two interlocutors interacting to solve various planning tasks for scheduling train routes and arranging railroad freight. One user acts the role of a planning assistant system and the other user acts as the coordinator. This was not done in a Wizard-of-Oz fashion, and as such is not considered a Human-Machine corpus. 34 different interlocutors were asked to complete 20 different tasks such as: "Determine the maximum number of boxcars of oranges that you could get to Bath by 7 AM tomorrow morning. It is now 12 midnight." The person playing the role of the planning assistant was provided with access to information that is needed to solve the task. Also included in the dataset is the information available to both users, the length of dialogue, and the speaker and 'system' interlocutor identities.

The **Verbmobil Corpus** (Burger et al., 2000) is a multilingual corpus consisting of English, German, and Japanese dialogues collected for the purposes of training and testing the Verbmobil project system. The system was a designed for speech-to-speech machine translation tasks. Dialogues were recorded in a variety of conditions and settings with room microphones, telephones, or close microphones, and were subsequently transcribed. Users were tasked with planning and scheduling an appointment throughout the course of the dialogue. Note that while there have been several versions of the Verbmobil corpora released, we refer to the entire collection here as described in (Burger et al., 2000). Dialogue acts were annotated in a subset of the corpus (1,505 mixed dialogues in German, English and Japanese). 76,210 acts were annotated with 32 possible categories of dialogue acts Alexandersson et al. (2000).[5]

Persuasion and Debates Another theme recurring among constrained spoken corpora is the appearance of persuasion or debate tasks. These can involve general debates on a topic or tasking a specific interlocutor to try to convince another interlocutor of some opinion or topic. Generally,

5. Note, this information and further facts about the Verbmobil project and corpus can be found here: http://verbmobil.dfki.de/facts.html

these datasets record the outcome of how convinced the audience is of the argument at the end of the dialogue or debate.

The **Green Persuasive Dataset** (Douglas-Cowie et al., 2007) was recorded in 2007 to provide data for the HUMAINE project, whose goal is to develop interfaces that can register and respond to emotion. In the dataset, a persuader with strong pro-environmental ('pro-green') feelings tries to convince persuadees to consider adopting more green lifestyles; these interactions are in the form of dialogues. It contains 8 long dialogues, totalling about 30 minutes each. Since the persuadees often either disagree or agree strongly with the persuaders points, this would be good corpus for studying social signs of (dis)-agreement between two people.

The **MAHNOB Mimicry Database** (Sun et al., 2011) contains 11 hours of recordings, split over 54 sessions between 60 people engaged either in a socio-political discussion or negotiating a tenancy agreement. This dataset consists of a set of fully synchronized audio-visual recordings of natural dyadic (one-on-one) interactions. It is one of several dialogue corpora that provide multi-modal data for analyzing human behaviour during conversations. Such corpora often consist of auditory, visual, and written transcriptions of the dialogues. Here, only audio-visual recordings are provided. The purpose of the dataset was to analyze mimicry (i.e. when one participant mimics the verbal and nonverbal expressions of their counterpart). The authors provide some benchmark video classification models to this effect.

The **Intelligence Squared Debate Dataset** (Zhang et al., 2016) covers the "Intelligence Squared" Oxford-style debates taking place between 2006 and 2015. The topics of the debates vary across the dataset, but are constrained within the context of each debate. Speakers are labeled and the full transcript of the debate is provided. Furthermore, the outcome of the debate is provided (how many of the audience members were for the given proposal or against, before and after the debate).

QA or Information Retrieval There are several corpora which feature direct question-and-answering sessions. These may involve general QA, such as in a press conference, or more task-specific lines of questioning to retrieve a specific set of information.

The **Corpus of Professional Spoken American English** (CPSAE) (Barlow, 2000) was constructed using a selection of transcripts of interactions occurring in professional settings. The corpus contains two million words involving over 400 speakers, recorded between 1994 and 1998. The CPASE has two main components. The first is a collection of transcripts (0.9 million words) of White House press conferences, which contains almost exclusively question and answer sessions, with some policy statements by politicians. The second component consists of transcripts (1.1 million words) of faculty meetings and committee meetings related to national tests that involve statements, discussions, and questions. The creation of the corpus was motivated by the desire to understand and model formal uses of the English language.

As previously mentioned, the Dialog State Tracking Challenge (DSTC) consists of a series of datasets evaluated using a 'state tracking' or 'slot filling' metric. While the first 3 installments of this challenge had conversations between a human participant and a computer, **DSTC4** (Kim et al., 2015) contains dialogues between humans. In particular, this dataset has 35 conversations with 21 hours of interactions between tourists and tour guides over Skype, discussing information on hotels, flights, and car rentals. Due to the small size of the dataset, researchers were encouraged to use transfer learning using other DSTC datasets to improve state tracking performance. This same training set is used for **DSTC5** (Kim et al., 2016) as well. However, the goal of DSTC5 is to study multi-lingual speech-act prediction, and therefore it combines the DSTC4 dialogues plus a set of equivalent Chinese dialogues; evaluation is done on a holdout set of Chinese dialogues.

Miscellaneous Lastly, there are several corpora which do not fall into any of the aforementioned categories, involving a range of tasks and situations.

The **IDIAP Wolf Corpus** (Hung and Chittaranjan, 2010) is an audio-visual corpus containing natural conversational data of volunteers who took part in an adversarial role-playing game called 'Werewolf'. Four groups of 8 to 12 people were recorded using headset microphones and synchronized video cameras, resulting in over 7 hours of conversational data. The novelty of this dataset is that the roles of other players are unknown to game participants, and some of the roles are deceptive in nature. Thus, there is a significant amount of lying that occurs during the game. Although specific instances of lying are not annotated, each speaker is labeled with their role in the game. In a dialogue setting, this could be useful for analyzing the differences in language when deception is being used.

The **SEMAINE Corpus** (McKeown et al., 2010) consists of 100 'emotionally coloured' conversations. Participants held conversations with an operator who adopted various roles designed to evoke emotional reactions. These conversations were recorded with synchronous video and audio devices. Importantly, the operators' responses were stock phrases that were independent of the content of the user's utterances, and only dependent on the user's emotional state. This corpus motivates building dialogue systems with affective and emotional intelligence abilities, since the corpus does not exhibit the natural language understanding that normally occurs between human interlocutors.

The **Loqui Human-Human Dialogue Corpus** (Passonneau and Sachar, 2014) consists of annotated transcriptions of telephone interactions between patrons and librarians at New York City's Andrew Heiskell Braille & Talking Book Library in 2006. It stands out as it has annotated discussion topics, question-answer pair links (adjacency pairs), dialogue acts, and frames (discourse units).

Similarly, the **The ICSI Meeting Recorder Dialog Act (MRDA) Corpus** (Shriberg et al., 2004) has annotated dialogue acts, question-answer pair links (adjacency pairs), and dialogue hot spots.[6] It consists of transcribed recordings of 75 ICSI meetings on several classes of topics including: the ICSI meeting recorder project itself, automatic speech recognition, natural language processing and neural theories of language, and discussions with the annotators for the project.

6. For more information on dialogue hot spots and how they relate to dialogue acts, see (Wrede and Shriberg, 2003).

Name	Topics	Total # of dialogues	Total # of words	Total length	Description
HCRC Map Task Corpus (Anderson et al., 1991)	Map-Reproducing Task	128	147k	18hrs	Dialogues from HLAP Task in which speakers must collaborate verbally to reproduce on one participants map a route printed on the others.
The Walking Around Corpus (Brennan et al., 2013)	Location Finding Task	36	300k*	33hrs	People collaborating over telephone to find certain locations.
Green Persuasive Database (Douglas-Cowie et al., 2007)	Lifestyle	8	35k*	4hrs	A persuader with (genuinely) strong pro-green feelings tries to convince persuadees to consider adopting more green lifestyles.
Intelligence Squared Debates (Zhang et al., 2016)	Debates	108	1.8M	200hrs*	Various topics in Oxford-style debates, each constrained to one subject. Audience opinions provided pre- and post-debates.
The Corpus of Professional Spoken American English (Barlow, 2000)	Politics, Education	200	2M	220hrs*	Interactions from faculty meetings and White House press conferences.
MAHNOB Mimicry Database (Sun et al., 2011)	Politics, Games	54	100k*	11hrs	Two experiments: a discussion on a political topic, and a role-playing game.
The IDIAP Wolf Corpus (Hung and Chittaranjan, 2010)	Role-Playing Game	15	60k*	7hrs	A recording of Werewolf role-playing game with annotations related to game progress.
SEMAINE corpus (McKeown et al., 2010)	Emotional Conversations	100	450k*	50hrs	Users were recorded while holding conversations with an operator who adopts roles designed to evoke emotional reactions.
DSTC4/DSTC5 Corpora (Kim et al., 2015, 2016)	Tourist	35	273k	21hrs	Tourist information exchange over Skype.
Loqui Dialogue Corpus (Passonneau and Sachar, 2014)	Library Inquiries	82	21K	140*	Telephone interactions between librarians and patrons. Annotated dialogue acts, discussion topics, frames (discourse units), question-answer pairs.
MRDA Corpus (Shriberg et al., 2004)	ICSI Meetings	75	11K*	72hrs	Recordings of ICSI meetings. Topics include: the corpus project itself, automatic speech recognition, natural language processing and theories of language. Dialogue acts, question-answer pairs, and hot spots.
TRAINS 93 Dialogues Corpus (Heeman and Allen, 1995)	Railroad Freight Route Planning	98	55K	6.5hrs	Collaborative planning of railroad freight routes.
Verbmobil Corpus (Burger et al., 2000)	Appointment Scheduling	726	270K	38Hrs	Spontaneous speech data collected for the Verbmobil project. Full corpus is in English, German, and Japanese. We only show English statistics.

Table 3: Human-human constrained spoken dialogue datasets. Starred (*) numbers are estimates based on the average rate of English speech from the National Center for Voice and Speech (www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html).

4.2.3 SCRIPTED CORPORA

A final category of spoken dialogue consists of conversations that have been pre-scripted for the purpose of being spoken later. We refer to datasets containing such conversations as 'scripted corpora'. As discussed in Subsection 3.5, these datasets are distinct from spontaneous human-human conversations, as they inevitably contain fewer 'filler' words and expressions that are common in spoken dialogue. However, they should not be confused with human-human written dialogues, as they are intended to sound like natural spoken conversations when read aloud by the participants. Furthermore, most of the works here are fictional — a distinction made in Subsection 3.5. As such, these scripted dialogues are required to be dramatic, as they are generally sourced from movies or TV shows.

There exist multiple scripted corpora based on movies and TV series. These can be sub-divided into two categories: corpora that provide the actual scripts (i.e. the movie script or TV series script) where each utterance is tagged with the appropriate speaker, and those that only contain subtitles and consecutive utterances are not divided or labeled in any way. It is always preferable to have the speaker labels, but there is significantly more unlabeled subtitle data available, and both sources of information can be leveraged to build a dialogue system.

The **Movie DiC Corpus** (Banchs, 2012) is an example of the former case—it contains about 130,000 dialogues and 6 million words from movie scripts extracted from the Internet Movie Script Data Collection[7], carefully selected to cover a wide range of genres. These dialogues also come with context descriptions, as written in the script. One derivation based on this corpus is the **Movie Triples Dataset** (Serban et al., 2016). There is also the **American Film Scripts Corpus** and **Film Scripts Online Corpus** which form the **Film Scripts Online Series Corpus**, which can be purchased.[8] The latter consists of a mix of British and American film scripts, while the former consists of solely American films.

The majority of these datasets consist of raw scripts, which are not guaranteed to portray conversations between only two people. The dataset collected by Nio et al. (2014), which we refer to as the **Filtered Movie Script Corpus**, takes over 1 million utterance-response pairs from web-based script resources and filters them down to 86,000 such pairs. The filtering method limits the extracted utterances to X-Y-X triples, where X is spoken by one actor and Y by another, and each of the utterances share some semantic similarity. These triples are then decomposed into X-Y and Y-X pairs. Such filtering largely removes conversations with more than two speakers, which could be useful in some applications. Particularly, the filtering method helps to retain semantic context in the dialogue and keeps a back-and-forth conversational flow that is desired in training many dialogue systems.

The **Cornell Movie-Dialogue Corpus** (Danescu-Niculescu-Mizil and Lee, 2011) also has short conversations extracted from movie scripts. The distinguishing feature of this dataset is the amount of metadata available for each conversation: this includes movie metadata such as genre, release year, and IMDB rating, as well as character metadata such as gender and position on movie credits. Although this corpus contains 220,000 dialogue excerpts, it only contains 300,000 utterances; thus, many of the excerpts contain a single utterance.

The **Corpus of American Soap Operas** (Davies, 2012b) contains 100 million words in more than 22,000 transcripts of ten American TV-series soap operas from 2001 and 2012. Because it is based on soap operas it is qualitatively different from the Movie Dic Corpus, which contains movies

7. http://www.imsdb.com
8. http://alexanderstreet.com/products/film-scripts-online-series

in the action and horror genres. The corpus was collected to provide insights into colloquial American speech, as the vocabulary usage is quite different from the British National Corpus (Davies, 2012a). Unfortunately, this corpus does not come with speaker labels.

Another corpus consisting of dialogues from TV shows is the **TVD Corpus** (Roy et al., 2014). This dataset consists of 191 movie transcripts from the comedy show *The Big Bang Theory*, and the drama show *Game of Thrones*, along with crowd-sourced text descriptions (brief episode summaries, longer episode outlines) and various types of metadata (speakers, shots, scenes). Text alignment algorithms are used to link descriptions and metadata to the appropriate sections of each script. For example, one might align an event description with all the utterances associated with that event in order to develop algorithms for locating specific events from raw dialogue, such as 'person X tries to convince person Y'.

Some work has been done in order to analyze character style from movie scripts. This is aided by a dataset collected by Walker et al. (2012a) that we refer to as the **Character Style from Film Corpus**. This corpus was collected from the IMSDb archive, and is annotated for linguistic structures and character archetypes. Features, such as the sentiment behind the utterances, are automatically extracted and used to derive models of the characters in order to generate new utterances similar in style to those spoken by the character. Thus, this dataset could be useful for building dialogue personalization models.

There are two primary movie subtitle datasets: the **OpenSubtitles** (Tiedemann, 2012) and the **SubTle Corpus** (Ameixa and Coheur, 2013). Both corpora are based on the OpenSubtitles website.[9] The OpenSubtitles dataset is a giant collection of movie subtitles, containing over 1 billion words, whereas SubTle Corpus has been pre-processed in order to extract interaction-response pairs that can help dialogue systems deal with out-of-domain (OOD) interactions.

The **Corpus of English Dialogues 1560-1760 (CED)** (Kytö and Walker, 2006) compiles dialogues from the mid-16th century until the mid-18th century. The sources vary from real trial transcripts to fiction dialogues. Due to the scripted nature of fictional dialogues and the fact that the majority of the corpus consists of fictional dialogue, we classify it here as such. The corpus is composed as follows: trial proceedings (285,660 words), witness depositions (172,940 words), drama comedy works (238,590 words), didactic works (236,640 words), prose fiction (223,890 words), and miscellaneous (25,970 words).

9. http://www.opensubtitles.org

Name	Topics	Total # of utterances	Total # of dialogues	Total # of works	Total # of words	Description
Movie-DiC (Banchs, 2012)	Movie dialogues	764k	132K	753	6M	Movie scripts of American films.
Movie-Triples (Serban et al., 2016)	Movie dialogues	736k	245K	614	13M	Triples of utterances which are filtered to come from X-Y-X triples.
Film Scripts Online Series	Movie scripts	1M*	263K†	1,500	16M*	Two subsets of scripts (1000 American films and 500 mixed). (British/American films).
Cornell Movie-Dialogue Corpus (Danescu-Niculescu-Mizil and Lee, 2011)	Movie dialogues	305K	220K	617	9M*	Short conversations from film scripts, annotated with character metadata.
Filtered Movie Script Corpus (Nio et al., 2014)	Movie dialogues	173k	87K	1,786	2M*	Triples of utterances which are filtered to come from X-Y-X triples.
American Soap Opera Corpus (Davies, 2012b)	TV show scripts	10M*	1.2M†	22,000	100M	Transcripts of American soap operas.
TVD Corpus (Roy et al., 2014)	TV show scripts	60k*	10K†	191	600k*	TV scripts from a comedy (Big Bang Theory) and drama (Game of Thrones) show.
Character Style from Film Corpus (Walker et al., 2012a)	Movie scripts	664k	151K	862	9.6M	Scripts from IMSDb, annotated for linguistic structures and character archetypes.
SubTle Corpus (Ameixa and Coheur, 2013)	Movie subtitles	6.7M	3.35M	6,184	20M	Aligned interaction-response pairs from movie subtitles.
OpenSubtitles (Tiedemann, 2012)	Movie subtitles	140M*	36M†	207,907	1B	Movie subtitles which are not speaker-aligned.
CED (1560–1760) Corpus (Kytö and Walker, 2006)	Written Works & Trial Proceedings	–	–	177	1.2M	Various scripted fictional works from (1560–1760) as well as court trial proceedings.

Table 4: Human-human scripted dialogue datasets. Quantities denoted with (†) indicate estimates based on average number of dialogues per movie (Banchs, 2012) and the number of scripts or works in the corpus. Dialogues may not be explicitly separated in these datasets. TV show datasets were adjusted based on the ratio of average film runtime (112 minutes) to average TV show runtime (36 minutes). This data was scraped from the IMBD database (http://www.imdb.com/interfaces). (Starred (*) quantities are estimated based on the average number of words and utterances per film, and the average lengths of films and TV shows. Estimates derived from the Tameri Guide for Writers (http://www.tameri.com/format/wordcounts.html).

4.3 Human-Human Written Corpora

We proceed to survey corpora of conversations between humans in written form. As before, we sub-divide this section into *spontaneous* and *constrained* corpora, depending on whether there are restrictions on the topic of conversation. However, we make a further distinction between *forum*, *micro-blogging*, and *chat* corpora.

Forum corpora consist of conversations on forum-based websites such as Reddit[10] where users can make posts, and other users can make comments or replies to said post. In some cases, comments can be nested indefinitely, as users make replies to previous replies. Utterances in forum corpora tend to be longer, and there is no restriction on the number of participants in a discussion. On the other hand, conversations on micro-blogging websites such as Twitter[11] tend to have very short utterances as there is an upper bound on the number of characters permitted in each message. As a result, these tend to exhibit highly colloquial language with many abbreviations. The identifying feature of chat corpora is that the conversations take place in real-time between users. Thus, these conversations share more similarities with spoken dialogue between humans, such as common grounding phenomena.

4.3.1 SPONTANEOUS WRITTEN CORPORA

We begin with written corpora where the topic of conversation is not pre-specified. Such is the case for the **NPS Internet Chatroom Conversations Corpus** (Forsyth and Martell, 2007), which consists of 10,567 English utterances gathered from age-specific chat rooms of various online chat services from October and November of 2006. Each utterance is annotated with part-of-speech and dialogue act information; the correctness of these labels was verified manually. The NPS Internet Chatroom Conversations Corpus was one of the first corpora of computer-mediated communication (CMC), and it was intended for various NLP applications such as conversation thread topic detection, author profiling, entity identification, and social network analysis.

Several corpora of spontaneous micro-blogging conversations have been collected, such as the **Twitter Corpus** from Ritter et al. (2010), which contains 1.3 million post-reply pairs extracted from Twitter. The corpus was originally constructed to aid in the production of unsupervised approaches to modeling dialogue acts. Larger Twitter corpora have been collected. The **Twitter Triples Corpus** (Sordoni et al., 2015) is one such example, with a described original dataset of 127 million context-message-response triples, but only a small labeled subset of this corpus has been released. Specifically, the released labeled subset contains 4,232 pairs that scored an average of greater than 4 on the Likert-type scale by crowdsourced evaluators for quality of the response to the context-message pair. Similarly, a large micro-blogging dataset, the **Sina Weibo Corpus** (Shang et al., 2015), which contains 4.5 million post-reply pairs, has been collected and used in literature, but this resource has not yet been made publicly available. We do not include the Sina Weibo Corpus (and its derivatives) in the tables in this section, as they are not primarily in English.

The **Usenet Corpus** (Shaoul and Westbury, 2009) is a gigantic collection of public Usenet postings[12] containing over 7 billion words from October 2005 to January 2011. Usenet was a distributed discussion system established in 1980 where participants could post articles to one of 47,860 'newsgroup' categories. It is seen as the precursor to many current Internet forums. The

10. `http://www.reddit.com`
11. `http://www.twitter.com`
12. `http://www.usenet.net`

corpus derived from these posts has been used for research in collaborative filtering (Konstan et al., 1997) and role detection (Fisher et al., 2006).

The **NUS SMS Corpus** (Chen and Kan, 2013) consists of conversations carried out over mobile phone SMS messages between two users. While the original purpose of the dataset was to improve predictive text entry when mobile phones still mapped multiple letters to a single number, aided by video and timing analysis of users entering their messages it could equally be used for analysis of informal dialogue. It is worth noting that the corpus does not consist of dialogues, but rather single SMS messages. SMS messages are similar in style to Twitter, in that they use many abbreviations and acronyms.

The **DailyDialog Dataset** (Li et al., 2017) consists of conversations crawled from websites which teach English through dialogue. These dialogues consist of everyday conversations such as a customer looking for a product or conversing with a salesperson. Since the collected data comes from educational sources, the dialogues are well-defined and generally free of grammatical mistakes or abbreviations. Furthermore, the data is hand-labeled with emotions. This added labeling may provide a useful complementary signal in training dialogue systems — for example as a latent variable for eliciting different types of emotion from a dialogue agent.

Currently, one of the most popular forum-based websites is Reddit where users can create discussions and post comments in various sub-forums called ‘subreddits’. Each subreddit addresses its own particular topic. Over 1.7 billion of these comments have been collected in the **Reddit Corpus**.[13] Each comment is labeled with the author, score (rating from other users), and position in the comment tree; the position is important as it determines which comment is being replied to. Researchers are just starting to investigate dialogue problems using this Reddit discussion corpus; its large size makes it a particularly interesting candidate for studying transfer learning. Additionally, researchers have used smaller collections of Reddit discussions for broad discourse classification (Schrading et al., 2015).

Some more specialized versions of the Reddit dataset have been curated. The **Reddit Domestic Abuse Corpus** (Schrading et al., 2015) consists of Reddit posts and comments taken from either subreddits specific to domestic abuse, or from subreddits representing casual conversations, advice, and general anxiety or anger. The motivation is to build classifiers that can detect occurrences of domestic abuse in other areas, which could provide insights into the prevalence and consequences of these situations. These conversations have been pre-processed with lower-casing, lemmatizing, and removal of stop words, and semantic role labels are provided.

13. See: `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`.

Name	Type	Topics	Avg. # of turns	Total # of dialogues	Total # of words	Description
NPS Chat Corpus (Forsyth and Martell, 2007)	Chat	Unrestricted	704	$15^{\dagger}$	100M	Posts from age-specific online chat rooms.
Twitter Corpus (Ritter et al., 2010)	Microblog	Unrestricted	2	1.3M	$125M^{\ddagger}$	Tweets and replies extracted from Twitter
Twitter Triple Corpus (Sordoni et al., 2015)	Microblog	Unrestricted	3	4,232	$65K^{\ddagger}$	A-B-A triples extracted from Twitter
UseNet Corpus (Shaoul and Westbury, 2009)	Microblog	Unrestricted	687	$47860^{\dagger}$	7B	UseNet forum postings
NUS SMS Corpus (Chen and Kan, 2013)	SMS messages	Unrestricted	18	3K	$580{,}668^{*\square}$	SMS messages collected between two users, with timing analysis.
Reddit	Forum	Unrestricted	–	–	–	1.7B comments across Reddit.
Reddit Domestic Abuse Corpus (Schrading et al., 2015)	Forum	Abuse help	17.53	21,133	19M-103M $^{\triangle}$	Reddit posts from either domestic abuse subreddits, or general chat.
Settlers of Catan (Afantenos et al., 2012)	Chat	Game terms	95	21	–	Conversations between players in the game 'Settlers of Catan.'
Cards Corpus (Djalali et al., 2012)	Chat	Game terms	38.1	1,266	282K	Conversations between players playing 'Cards world.'
Agreement in Wikipedia Talk Pages (Andreas et al., 2012)	Forum	Unrestricted	2	822	110K	LiveJournal and Wikipedia Discussions forum threads. Agreement type and level annotated.
Agreement by Create Debaters (Rosenthal and McKeown, 2015)	Forum	Unrestricted	2	10K	1.4M	Create Debate forum conversations. Annotated with type of agreement or disagreement.
Internet Argument Corpus (Walker et al., 2012b)	Forum	Politics	35.45	11K	73M	Debates about specific political or moral positions.
MPC Corpus (Shaikh et al., 2010)	Chat	Social tasks	520	14	58K	Conversations about general, political, and interview topics.
Ubuntu Dialogue Corpus (Lowe et al., 2015a)	Chat	Ubuntu Operating System	7.71	930K	100M	Dialogues extracted from Ubuntu chat stream on IRC.
Ubuntu Chat Corpus (Uthus and Aha, 2013)	Chat	Ubuntu Operating System	3381.6	$10665^{\dagger}$	$2B^{*\square}$	Chat stream scraped from IRC logs (no dialogues extracted).
Movie Dialog Dataset (Dodge et al., 2015)	Chat, QA & Recommendation	Movies	3.3	$3.1M^{\blacktriangledown}$	185M	For goal-driven dialogue systems. Includes movie metadata as knowledge triples.
DailyDialog (Li et al., 2017)	Chat	Daily Life	7.9	13K	1.5M	Conversations extracted from English language educational texts. Labeled with emotions.

Table 5: Human-human written dialogue datasets. Starred (*) quantities are computed using word counts based on spaces. Triangle ($^{\triangle}$) indicates lower and upper bounds computed using average words per utterance estimated on a similar Reddit corpus Schrading (2015). Square ($^{\square}$) indicates estimates based only on the English part of the corpus. Dialogues indicated by (†) are contiguous blocks of recorded conversation in a multi-participant chat. For UseNet, the average number of turns are calculated as the average number of posts collected per newsgroup. (‡) indicates an estimate based on a Twitter dataset of similar size and refers to tokens as well as words.

4.3.2 CONSTRAINED WRITTEN CORPORA

There are also several written corpora where users are limited in terms of topics of conversation. For example, the **Settlers of Catan Corpus** (Afantenos et al., 2012) contains logs of 40 games of 'Settlers of Catan', with about 80,000 total labeled utterances. The game is played with up to 4 players, and is predicated on trading certain goods between players. The goal of the game is to be the first player to achieve a pre-specified number of points. Therefore, the game is adversarial in nature, and can be used to analyze situations of strategic conversation where the agents have diverging motives.

Another corpus that deals with game playing is the **Cards Corpus** (Djalali et al., 2012), which consists of 1,266 transcripts of conversations between players playing a game in the 'Cards world'. This world is a simple 2-D environment where players collaborate to collect cards. The goal of the game is to collect six cards of a particular suit (cards in the environment are only visible to a player when they are near the location of that player), or to determine that this goal is impossible in the environment. The catch is that each player can only hold 3 cards, thus players must collaborate in order to achieve the goal. Further, each player's location is hidden to the other player, and there are a fixed number of non-chatting moves. Thus, players must use the chat to formulate a plan, rather than exhaustively exploring the environment themselves. The dataset has been further annotated by Potts (2012) to collect all locative question-answer pairs (i.e. all questions of the form "Where are you?").

The **Agreement by Create Debaters Corpus** (Rosenthal and McKeown, 2015), the **Agreement in Wikipedia Talk Pages Corpus** (Andreas et al., 2012) and the **Internet Argument Corpus** (Abbott et al., 2016) all cover dialogues with annotations measuring levels of agreement or disagreement in responses to posts in various media. The **Agreement by Create Debaters Corpus** and the **Agreement in Wikipedia Talk Pages Corpus** both are formatted in the same way. Post-reply pairs are annotated with whether they are in agreement or disagreement, as well as the type of agreement they are in if applicable (e.g. paraphrasing). The difference between the two corpora is the source: the former is collected from Create Debate forums and the latter from a mix of Wikipedia Discussion pages and LiveJournal postings. The **Internet Argument Corpus** (IAC) (Walker et al., 2012b) is a forum-based corpus with 390,000 posts on 11,000 discussion topics. Each topic is controversial in nature, including subjects such as evolution, gay marriage and climate change; users participate by sharing their opinions on one of these topics. Posts-reply pairs have been labeled as being either in agreement or disagreement, and sarcasm ratings are given to each post.

Another source of constrained text-based corpora are chat-room environments. Such a set-up forms the basis of the **MPC Corpus** (Shaikh et al., 2010), which consists of 14 multi-party dialogue sessions of approximately 90 minutes each. In some cases, discussion topics were constrained to be about certain political stances, or mock committees for choosing job candidates. An interesting feature is that different participants are given different roles — leader, disruptor, and consensus builder — with only a general outline of their goals in the conversation. Thus, this dataset could be used to model social phenomena such as agenda control, influence, and leadership in on-line interactions.

The largest written corpus with a constrained topic is the recently released **Ubuntu Dialogue Corpus** (Lowe et al., 2015a), which has almost 1 million dialogues of 3 turns or more, and 100 million words. It is related to the former **Ubuntu Chat Corpus** (Uthus and Aha, 2013). Both corpora were scraped from the Ubuntu IRC channel logs.[14] On this channel, users can log in and ask a question about a problem they are having with Ubuntu; these questions are answered by other users. Although the chat room allows everyone to chat with each other in a multi-party setting, the Ubuntu Dialogue Corpus uses a series of heuristics to disentangle it into dyadic dialogue. The technical nature and size of this corpus lends itself particularly well to applications in technical support.

Other corpora have been extracted from IRC chat logs. The **IRC Corpus** (Elsner and Charniak, 2008) contains approximately 50 hours of chat, with an estimated 20,000 utterances from the Linux channel on IRC, complete with the posting times. Therefore, this dataset consists of technical conversations similar to the Ubuntu Corpus, with the occasional social chat. The purpose of this dataset was to investigate approaches for conversation disentanglement; given a multi-party chat room, one attempts to recover the individual conversations of which it is composed. For this purpose, there are approximately 1,500 utterances with annotated ground-truth conversations.

More recent efforts have combined traditional conversational corpora with question answering and recommendation datasets in order to facilitate the construction of goal-driven dialogue systems. Such is the case for the **Movie Dialog Dataset** (Dodge et al., 2015). There are four tasks that the authors propose as a prerequisite for a working dialogue system: question answering, recommendation, question answering with recommendation, and casual conversation. The Movie Dialog dataset consists of four sub-datasets used for training models to complete these tasks: a QA dataset from the Open Movie Database (OMDb)[15] of 116k examples with accompanying movie and actor metadata in the form of knowledge triples; a recommendation dataset from MovieLens[16] with 110k users and 1M questions; a combined recommendation and QA dataset with 1M conversations of 6 turns each; and a discussion dataset from Reddit's movie subreddit. The former is evaluated using recall metrics in a manner similar to Lowe et al. (2015a). It should be noted that, other than the Reddit dataset, the dialogues in the sub-datasets are simulated QA pairs, where each response corresponds to a list of entities from the knowledge base.

5. Discussion

We now discuss a number of challenges and general methods related to the development and evaluation of data-driven dialogue systems. We highlight challenges relevant to working with large-scale datasets, colloquial language, spelling mistakes and acronyms, as well as missing and unobservable data. We also discuss methods to improve data-driven dialogue systems beyond a single corpus, such as transfer learning between datasets and the usage of external knowledge. Researchers and developers may consider applying these methods once they have settled on using one or several corpora. We also discuss user personalization, applicable in the case where there is rich information available for each user. Finally, we discuss different methods for evaluating data-driven dialogue systems, including corpus-based evaluation methods. Evaluating a data-driven dialogue system properly is critical for real-world deployments as well as for advancing state-of-the-art research, in which case reproducibility of methods and results is crucial.

14. http://irclogs.ubuntu.com
15. http://en.omdb.org
16. http://movielens.org

5.1 Challenges of Learning from Large Datasets

Recently, several of the larger dialogue datasets have been used to train data-driven dialogue systems; the Twitter Corpus (Ritter et al., 2010) and the Ubuntu Dialogue corpus (Lowe et al., 2015a) are two examples. In this section, we discuss the benefits and drawbacks of these datasets based on our experience using them for building data-driven models. Unlike the previous section, we now focus on highly relevant aspects and characteristics of these datasets specifically for learning in data-driven dialogue systems based on neural-network architectures.

5.1.1 THE TWITTER CORPUS

The Twitter Corpus consists of a series of conversations extracted from tweets. While the dataset is large and general-purpose, the micro-blogging nature of the source material leads to several drawbacks for building conversational dialogue agents. However, some of these drawbacks do not apply if the end goal is to build an agent that interacts with users on the Twitter platform.

The Twitter Corpus has an enormous amount of typos, slang, and abbreviations. Due to the 140-character limit in this dataset, tweets are often very short and compressed. In addition, users frequently use Twitter-specific devices such as hashtags. Unless one is building a dialogue agent specifically for Twitter, it is often not desirable to have a chatbot use hashtags and excessive abbreviations as it is not reflective of how humans converse in other environments. This also results in a significant increase in the word vocabulary required for dialogue systems trained at the word level. As such, it is not surprising that character-level models have shown promising results on Twitter (Dhingra et al., 2016).

Twitter conversations often contain various kinds of verbal role-playing and imaginative actions similar to stage directions in theater plays (e.g. instead of writing *"goodbye"*, a user might write *"*waves goodbye and leaves*"*). These conversations are very different from the majority of text-based chats. Therefore, dialogue models trained on this dataset are often able to provide interesting and accurate responses to contexts involving role-playing and imaginative actions (Serban et al., 2017d).

Another challenge is that Twitter conversations often rely on implicit context (e.g. they refer to recent public events outside the conversation). In order to learn effective responses for such conversations, a dialogue agent must infer the news event under discussion by referencing some form of external knowledge base. This would appear to be a particularly difficult task.

5.1.2 THE UBUNTU DIALOGUE CORPUS

The Ubuntu Dialogue Corpus is one of the largest, publicly available datasets containing technical support dialogues. Due to the commercial importance of such systems, the dataset has attracted significant attention.[17] Thus, the Ubuntu Dialogue Corpus presents opportunities for anyone to train large-scale data-driven technical support dialogue systems.

Despite this, there are several challenges when training data-driven dialogue models on the Ubuntu Dialogue Corpus due to the nature of the data. First, since the corpus comes from a multi-party IRC channel, it needs to be disentangled into separate dialogues. This notion of disentanglement in dialogue corpora — that is, given a multi-party dialogue, each utterance must be attributed

17. Most of the largest technical support datasets are based on commercial technical support channels, which are proprietary and never released to the public for privacy reasons.

to a conversational thread — has been investigated in several works Elsner and Charniak (2010, 2008). This disentanglement process is noisy, and errors inevitably arise. As a result, some cohesion can be lost and confusion introduced. The most frequent error is when a missing utterance in the dialogue is not picked up by the extraction procedure (i.e. an utterance from the original multi-party chat was not added to the disentangled dialogue). As a result, for a substantial amount of conversations, it is difficult to follow the topic. In particular, this means that some of the Next Utterance Classification (NUC) examples, where models must select the correct next response from a list of candidates, are either difficult or impossible for models to predict.

Another problem arises from the lack of annotations and labels. Since users try to solve their technical problems, it is perhaps best to build models under a goal-driven dialogue framework, where a dialogue system has to maximize the probability that it will solve the user's problem at the end of the conversation. However, there are no reward labels available. Thus, it is difficult to model the dataset in a goal-driven dialogue framework. Future work may alleviate this by constructing automatic methods of determining whether a user in a particular conversation solved their problem.

A particular challenge of the Ubuntu Dialogue Corpus is the large number of out-of-vocabulary words, including many technical words related to the Ubuntu operating system, such as commands, software packages, websites, etc. Since these words occur rarely in the dataset, it is difficult to learn their meaning directly from the dataset — for example, it is difficult to obtain meaningful distributed, real-valued vector representations for neural network-based dialogue models. This is further exacerbated by the large number of users who use different nomenclature, acronyms, and speaking styles, and the many typos in the dataset. Thus, the linguistic diversity of the corpus is large.

A final challenge of the dataset is the necessity for additional knowledge related to Ubuntu in order to accurately generate or predict the next response in a conversation. We hypothesize that this knowledge is crucial for a system trained on the Ubuntu Dialogue Corpus to be effective in practice, as often solutions to technical problems change over time as new versions of the operating system become available. Thus, an effective dialogue system must learn to combine up-to-date technical information with an understanding of natural language dialogue in order to solve the users' problems. We will discuss the use of external knowledge in more detail in Section 5.3.

While these challenges make it difficult to build data-driven dialogue systems, it also presents an important research opportunity. Current data-driven dialogue systems perform rather poorly in terms of generating utterances that are coherent and on-topic (Serban et al., 2017a). As such, there is significant room for improvement on these models.

5.2 Transfer Learning Between Datasets

While it is not always feasible to obtain large corpora for every new application, the use of other related datasets can effectively bootstrap the learning process. In several branches of machine learning, and in particular in deep learning, the use of related datasets for *pre-training* models is an effective method of scaling up to complex environments (Erhan et al., 2010; Kumar et al., 2015).

To build open-domain dialogue systems, it is arguably necessary to move beyond domain-specific datasets. Instead, like humans, dialogue systems may have to be trained on multiple data sources for solving multiple tasks. To leverage statistical efficiency, it may be necessary to first use unsupervised learning — as opposed to supervised learning or offline reinforcement learning, which typically only provide a sparse scalar feedback signal for each phrase or sequence of phrases — and then fine-tune models based on human feedback. Researchers have already proposed various ways of applying transfer learning to build data-driven dialogue systems, ranging from learning separate sub-components of the dialogue system (e.g. intent and dialogue act classification) to learning the entire dialogue system (e.g. in an unsupervised or reinforcement learning framework) using transfer learning (Fabbrizio et al., 2004; Forgues et al., 2014; Serban and Pineau, 2015; Serban et al., 2016; Lowe et al., 2015a; Vandyke et al., 2015; Wen et al., 2016; Gašić et al., 2016; Mo et al., 2016; Genevay and Laroche, 2016; Chen et al., 2016)

5.3 Incorporating External Knowledge

Another interesting research direction is the incorporation of external knowledge sources in order to inform the response to be generated. Using external information is of great importance to dialogues systems, particularly in the goal-driven setting. Even non-goal-driven dialogue systems designed to simply entertain the user could benefit from leveraging external information, such as current news articles or movie reviews, in order to better converse about real-world events. This may be particularly useful in data-sparse domains, when there is insufficient dialogue training data to reliably learn a response that is appropriate for each input utterance, or in domains that evolve quickly over time.

5.3.1 Structured External Knowledge

In traditional goal-driven dialogue systems (Levin and Pieraccini, 1997), where the goal is to provide information to the user, there is already extensive use of external knowledge sources. For example, in the Let's Go! dialogue system (Raux et al., 2005), the user requests information about various bus arrival and departure times. Thus, a critical input to the model is the actual bus schedule, which is used in order to generate the system's utterances. Another example is the dialogue system described by Nöth et al. (2004), which helps users find movie information by utilizing movie show times from different cinemas. Such examples are abundant both in the literature and in practice. Although these models make use of external knowledge, the knowledge sources in these cases are highly structured and are only used to place hard constraints on the possible states of an utterance to be generated. They are essentially contained in relational databases or structured ontologies, and are only used to provide a deterministic mapping from the dialogue states extracted from an input user utterance to the dialogue system state or the generated response.

Complementary to domain-specific databases and ontologies are the general natural language processing databases and tools. These include lexical databases such as WordNet (Miller, 1995), which contains lexical relationships between words for over a hundred thousand words, VerbNet (Schuler, 2005) which contains lexical relations between verbs, and FrameNet (Ruppenhofer et al., 2006), which contains 'word senses' for over ten thousand words along with examples of each word sense. In addition, there exist several natural language processing tools such as part of speech taggers, word category classifiers, word embedding models, named entity recognition models, co-reference resolution models, semantic role labeling models, semantic similarity models and sentiment analysis models (Manning and Schütze, 1999; Jurafsky and Martin, 2008; Mikolov et al., 2013; Gurevych and Strube, 2004; Lin and Walker, 2011b) that may be used by the Natural Language Understanding component to extract meaning from human utterances. Since these tools are typically built upon texts and annotations created by humans, using them inside a dialogue system can be interpreted as a form of structured transfer learning, where the relationships or labels learned from the original natural language processing corpus provide additional information to the dialogue system and improve generalization of the system.

5.3.2 UNSTRUCTURED EXTERNAL KNOWLEDGE

Complementary sources of information can be found in unstructured knowledge sources, such as online encyclopedias (Wikipedia (Denoyer and Gallinari, 2007)) as well as domain-specific sources (Lowe et al., 2015b). It is beyond the scope of this paper to review all possible ways that these unstructured knowledge sources have or could be used in conjunction with a data-driven dialogue system. However, we note that this is likely to be a fruitful research area.

5.4 Personalized dialogue agents

When conversing, humans often adapt to their interlocutor to facilitate understanding, and thus improve conversational efficiency and satisfaction. Attaining human-level performance with dialogue agents may well require personalization, i.e. models that are aware and capable of adapting to their interlocutor. Such capabilities could increase the effectiveness and naturalness of generated dialogues (Lucas et al., 2009; Su et al., 2013). We see personalization of dialogue systems as an important task, which so far has not received much attention. There has been initial efforts on user-specific models which could be adapted to work in combination with the dialogue models presented in this survey (Lucas et al., 2009; Lin and Walker, 2011a; Pargellis et al., 2004). There has also been interesting work on character modeling in movies (Walker et al., 2011; Li et al., 2016; Mo et al., 2016). There is significant potential to learn user models as part of dialogue models. The large datasets presented in this paper, some of which provide multiple dialogues per user, may enable the development of such models.

5.5 Evaluation metrics

One of the most challenging aspects of constructing dialogue systems lies in their evaluation. While the end goal is to deploy the dialogue system in an application setting and receive real human feedback, getting to this stage is time consuming and expensive. Often it is also necessary to optimize performance on a pseudo-performance metric prior to release. This is particularly true if a dialogue model has many hyper-parameters to be optimized — it is infeasible to run user experiments for every parameter setting in a grid search. Although crowdsourcing platforms, such as Amazon Mechanical Turk, can be used for some user testing (Jurcıcek et al., 2011), evaluations using paid subjects can also lead to biased results (Young et al., 2013). Ideally, we would have some automated metrics for calculating a score for each model, and only involve human evaluators once the best model has been chosen with reasonable confidence.

In non-goal-driven dialogue systems researchers have focused mainly on the output of the response generation module. Evaluation of such non-goal-driven dialogue systems can be traced back to the Turing test (Turing, 1950), where human judges communicate with both computer programs and other humans over a chat terminal without knowing the other party's true identity. The judge's goal is to identify the humans and computer programs under the assumption that a program indistinguishable from a real human being must be intelligent. However, this setup has been criticized extensively with numerous researchers proposing alternative evaluation procedures (Cohen, 2005). More recently, researchers have turned to analyzing the collected dialogues produced after they are finished (Galley et al., 2015; Pietquin and Hastie, 2013; Shawar and Atwell, 2007a; Schatzmann et al., 2005).

Even when human evaluators are available, it is often difficult to choose a set of informative and consistent criteria that can be used to judge an utterance generated by a dialogue system. For example, one might ask the evaluator to rate the f on vague notions such as 'appropriateness' and 'naturalness', or to try to differentiate between utterances generated by the system and those generated by actual humans (Vinyals and Le, 2015). Schatzmann et al. (2005) suggest two aspects that need to be evaluated for all response generation systems (as well as user simulation models): 1) if the model can generate human-like output, and 2) if the model can reproduce the variety of user behaviour found in corpus. But we lack a definitive framework for such evaluations.

We complete this discussion by summarizing different approaches to the automatic evaluation problem as they relate to these objectives.

5.5.1 AUTOMATIC EVALUATION METRICS FOR GOAL-DRIVEN DIALOGUE SYSTEMS

User evaluation of goal-driven dialogue systems typically focuses on goal-related performance criteria, such as goal completion rate, dialogue length, and user satisfaction (Walker et al., 1997; Schatzmann et al., 2005). These were originally evaluated by human users interacting with the dialogue system. Recently, researchers have also begun to use third-party annotators for evaluating recorded dialogues (Yang et al., 2010). Due to their simplicity, the vast majority of hand-crafted task-oriented dialogue systems have been solely evaluated in this way. However, when using machine learning algorithms to train on large-scale corpora, optimization criteria are required. The challenge with evaluating goal-driven dialogue systems without human intervention is that the process necessarily requires multiple steps — it is difficult to determine if a task has been solved from a single utterance-response pair from a conversation. Thus, synthetic data is often generated by a *user simulator* (Eckert et al., 1997; Schatzmann et al., 2007; Jung et al., 2009; Georgila et al., 2006; Pietquin and Hastie, 2013). Given a sufficiently accurate user simulation model, an interaction between the dialogue system and the user can be simulated from which it is possible to deduce the desired metrics, such as goal completion rate. Significant effort has been made to render the simulated data as realistic as possible, by modeling user intentions. Evaluation of such simulation methods has already been conducted (Schatzmann et al., 2005). However, generating realistic user simulation models remains an open problem.

5.5.2 AUTOMATIC EVALUATION METRICS FOR NON-GOAL-DRIVEN DIALOGUE SYSTEMS

Evaluation of non-goal-driven dialogue systems, whether by automatic means or user studies, remains a difficult challenge.

Word Overlap Metrics. One approach is to borrow evaluation metrics from other NLP tasks such as machine translation, which uses BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) scores. These metrics have been used to compare responses generated by a learned dialogue strategy to the actual next utterance in the conversation, conditioned on a dialogue context (Sordoni et al., 2015). However, BLEU scores have been shown not to correlate with human judgment for assessing dialogue response generation (Liu et al., 2016). There are several issues to consider: given the context of a conversation, there often exists a large number of possible responses that 'fit' into the dialogue. Thus, the response generated by a dialogue system could be entirely reasonable, yet it may have no words in common with the actual next utterance. In this case, the BLEU score would be very low, but would not accurately reflect the strength of the model. Indeed, even humans who are tasked with predicting the next utterance of a conversation achieve relatively low BLEU scores (Sordoni et al., 2015). Although the METEOR metric takes into account synonyms and morphological variants of words in the candidate response, it still suffers from the aforementioned problems. In a sense, these measurements only satisfy one direction of Schatzmann's criteria (Schatzmann et al., 2005): high BLEU and METEOR scores imply that the model is generating human-like output, but the model may still not reproduce the variety of user behaviour found in corpus. Furthermore, such metrics will only accurately reflect the performance of the dialogue system if given a large number of candidate responses for each given context.

Next Utterance Classification. Alternatively, one can narrow the number of possible responses to a small, pre-defined list, and ask the model to select the most appropriate response from this list. The list includes the actual next response of the conversation (the desired prediction), and the other entries (false positives) are sampled from elsewhere in the corpus (Lowe et al., 2016, 2015a). This *next utterance classification* (NUC) task is derived from recall and precision metrics typical of information-retrieval-based approaches. There are several attractive properties of this task: it is easy to interpret, and its difficulty can be adjusted by changing the number of false responses. However, there are drawbacks. Since the other candidate answers are sampled from elsewhere in the corpus, there is a chance that these also represent reasonable responses given the context. This can be alleviated to some extent by reporting Recall@k measures, i.e. whether the correct response is found in the k responses with the highest rankings according to the model. Although current models evaluated using NUC are trained explicitly to maximize the performance on a related metric (cross-entropy between context-response pairs (Lowe et al., 2015a; Kadlec et al., 2015)), precision and recall could also be used to evaluate a probabilistic generative model trained to outputs full utterances.

Word Perplexity. Another metric proposed to evaluate probabilistic language models (Bengio et al., 2003; Mikolov et al., 2010) that has seen significant recent use for evaluating end-to-end dialogue systems is *word perplexity* (Pietquin and Hastie, 2013; Serban et al., 2016). Perplexity explicitly measures the probability that the model will generate the ground truth next utterance given some context of the conversation. This is particularly appealing for dialogue, as the distribution over words in the next utterance can be highly multi-modal (i.e. many possible responses). A re-weighted perplexity metric has also been proposed where stop words, punctuation, and end-of-utterance tokens are ignored to focus on the semantic content of the phrase (Serban et al., 2016). Both word perplexity, as well as utterance-level recall and precision outlined above, satisfy Schatzmann's evaluation criteria, since scoring high on these would require the model to produce human-like output and to reproduce most types of conversations in the corpus.

Response Diversity. Recent non-goal-driven dialogue systems based on neural networks have had problems generating diverse responses (Serban et al., 2016). (Li et al., 2015) recently introduced two new metrics, *distinct-1* and *distinct-2*, which respectively measure the number of distinct unigrams and bigrams of the generated responses. Although these fail to satisfy either of Schatzmann's criteria, they may still be useful in combination with other metrics, such as BLEU, NUC or word perplexity.

6. Conclusion

This paper provides an extensive survey of currently available datasets suitable for research, development, and evaluation of data-driven dialogue systems. We categorize these corpora along several dimensions depending on whether the dataset is written or spoken, between human interlocutors or human-machine conversations, and constrained in topic or more free-form. We collect statistics for these datasets and present them in tables in Section 4, and provide an open-source GitHub repository where these datasets can be viewed and pull requests can be made to add new datasets: `https://github.com/Breakend/DialogDatasets`.

There is broad coverage of existing datasets along most of the dimensions we consider. However, the vast majority of the available datasets contain at most thousands of dialogues. This presents some challenges to the training of large-scale end-to-end models, such as neural networks, on general purpose domains. Neural networks can be applied to narrow domains, such as restaurant recommendation, with relatively little data (Wen et al., 2017). However, as the nature of interactions becomes more open and the number of topics grows, the sample complexity and with it the required dataset size increases. To obtain reasonable results in such a setting, neural network practitioners have resorted to training neural network models on datasets with hundreds of thousands to millions of dialogues: the Twitter Corpus (Ritter et al., 2010; Sordoni et al., 2015), Reddit, the Ubuntu Dialogue Corpus (Lowe et al., 2015a), and various movie subtitle datasets such as SubTle, OpenSubtitles, Movie-DiC, and the Movie Dialogue Dataset (Ameixa and Coheur, 2013; Tiedemann, 2012; Banchs, 2012; Dodge et al., 2015). While the conversation topics in these datasets often vary considerably, the nature of the datasets themselves are fairly fixed in the form of *informal written* dialogues between humans. This is the case for movie scripts, forum posts, and micro-blogging platforms. Learning only from these sources will bias dialogue systems towards certain kinds of interactions and behaviours; for example, written corpora usually have a specific turn-taking structure that is different from spoken conversation, and they may encode biases against certain groups or populations (Henderson et al., 2017). If we want dialogue systems to speak in a more natural way, similar to spoken human-human conversation, emphasis should be placed on collecting large-scale spoken dialogue corpora to train the next generation of dialogue systems. There is also a lack of large-scale multi-modal datasets, which may be crucial towards grounding the language learned by our dialogue agents in human-like experience.

We outline different approaches for overcoming the dearth of very large dialogue datasets. Transfer learning appears to be a particularly promising avenue for future dialogue research. While many individual datasets presented in Section 4 are only thousands of dialogues, summed together they represent a significant resource covering a wide range of topics. It would seem highly advantageous if methods are developed that enable dialogue systems to learn across all of these resources. There is also conceivable transfer that could be gained through learning on non-dialogue text corpora, such as Wikipedia. We further discuss research directions for building data-driven dialogue systems, including incorporating external knowledge and personalizing dialogue agents. We discuss several challenges associated with training large-scale dialogue models on two popular dialogue corpora: the Twitter Corpus and the Ubuntu Dialogue Corpus, based on our own experience.

Finally, we discuss automatic evaluation metrics used to train data-driven dialogue systems on the datasets previously mentioned. Having an automatic metric that correlates highly with human judgment of dialogue quality is very important. Even with significant dialogue data, poor evaluation metrics mean that it is difficult to compare the quality of dialogue systems trained on these datasets, and progress as a field becomes difficult. At present, there exists no silver bullet for automatic evaluation, implying that a set of diverse metrics be used together to obtain an accurate measure of performance. The final arbiter is, of course, human judgments. However, these can be expensive to obtain on a large scale, in particular for publicly-funded research laboratories. As a dialogue community, we should strive towards releasing more large-scale dialogue datasets and producing standardized evaluation metrics, to make dialogue research as inclusive as possible for all who work in the field.

Acknowledgements

The authors gratefully acknowledge financial support by the Samsung Advanced Institute of Technology (SAIT), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs, the Canadian Institute for Advanced Research (CIFAR) and Compute Canada. The second author is funded by a Vanier Graduate Scholarship. Early versions of the manuscript benefited greatly from the proofreading of Melanie Lyman-Abramovitch, and later versions were extensively revised by Genevieve Fried and Nicolas Angelard-Gontier. The authors also thank Nissan Pow, Michael Noseworthy, Chia-Wei Liu, Gabriel Forgues, Alessandro Sordoni, Yoshua Bengio and Aaron Courville for helpful discussions.

References

C. Khatri A. Venkatesh R. Gabriel Q. Li J. Nunn B. Hedayatnia M. Heng A. Nagar E. King K. Bland A. Wartick Y. Pan H. Song S. Jayadevan G. Hwang A. Pettigrue A. Ram, R. Prasad. Conversational ai: The science behind the alexa prize. In *Alexa Prize Proceedings*, 2017.

B. Aarts and S. A. Wallis. The diachronic corpus of present-day spoken english (DCPSE), 2006.

R. Abbott, B. Ecker, P. Anand, and M. Walker. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Language Resources and Evaluation Conference, LREC2016*, 2016.

S. Afantenos, N. Asher, F. Benamara, A. Cadilhac, Cédric Dégremont, P. Denis, M. Guhe, S. Keizer, A. Lascarides, O. Lemon, et al. Developing a corpus of strategic conversation in the settlers of catan. In *SeineDial 2012-The 16th workshop on the semantics and pragmatics of dialogue*, 2012.

H. Ai, A. Raux, D. Bohus, M. Eskenazi, and D. J. Litman. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2007.

R. Akker and D. Traum. A comparison of addressee detection methods for multiparty conversations. In *Workshop on the Semantics and Pragmatics of Dialogue*, 2009.

Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. M., and K. Yamada. Translating with scarce resources. In *AAAI*, 2000.

J. Alexandersson, R. Engel, M. Kipp, S. Koch, U. Küssner, N. Reithinger, and M. Stede. Modeling negotiation dialogs. In *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 441–451. Springer, 2000.

D. Ameixa and L. Coheur. From subtitles to human interactions: introducing the subtle corpus. Technical report, Tech. rep., 2013.

A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, et al. The HCRC map task corpus. *Language and speech*, 34(4):351–366, 1991.

J. Andreas, S. Rosenthal, and K. McKeown. Annotating agreement and disagreement in threaded discussion. In *LREC*, pages 818–822, 2012.

L. E. Asri, J. He, and K. Suleman. A sequence-to-sequence model for user simulation in spoken dialogue systems. *arXiv preprint arXiv:1607.00070*, 2016.

A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven. Cardiff conversation database (CCDb): A database of natural dyadic conversations. In *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Conference on*, pages 277–282, 2013.

H. Aust, M. Oerder, F. Seide, and V. Steinbiss. The philips automatic train timetable information system. *Speech Communication*, 17(3):249–262, 1995.

R. E. Banchs. Movie-DiC: a movie dialogue corpus for research and development. In *Association for Computational Linguistics: Short Papers*, 2012.

R. E. Banchs and H. Li. IRIS: a chat-oriented dialogue system based on the vector space model. In *Association for Computational Linguistics, System Demonstrations*, 2012.

S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Association for Computational Linguistics, Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

M. Barlow. Corpus of spoken, professional american-english, 2000.

J. Beare and B. Scott. The spoken corpus of the survey of english dialects: language variation and oral history. In *Proceedings of ALLC/ACH*, 1999.

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

Y. Bengio, I. Goodfellow, and A. Courville. Deep learning. *An MIT Press book in preparation. Draft chapters available at http://www. iro. umontreal. ca/ bengioy/dlbook*, 2014.

C. Bennett and A. I Rudnicky. The carnegie mellon communicator corpus, 2002.

D. Biber and E. Finegan. An initial typology of english text types. *Corpus linguistics II: New studies in the analysis and exploitation of computer corpora*, pages 19–46, 1986.

D. Biber and E. Finegan. Diachronic relations among speech-based and written registers in english. *Variation in English: multi-dimensional studies*, pages 66–83, 2001.

D. Bohus and A. I Rudnicky. Sorry, I didnt catch that! In *Recent Trends in Discourse and Dialogue*, pages 123–154. Springer, 2008.

S. E. Brennan, K. S. Schuhmann, and K. M. Batres. Entrainment on the move and in the lab: The walking around corpus. In *Conference of the Cognitive Science Society*, 2013.

G. Brown, A. Anderson, R. Shillcock, and G. Yule. Teaching talk. *Cambridge: CUP*, 1984.

S. Burger, K. Weilhammer, F. Schiel, and H. G. Tillmann. Verbmobil data collection and annotation. In *Verbmobil: Foundations of speech-to-speech translation*, pages 537–549. Springer, 2000.

J. E. Cahn and S. E. Brennan. A psychological model of grounding and repair in dialog. In *AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

A. Canavan and G. Zipperlen. Callfriend american english-non-southern dialect. *Linguistic Data Consortium*, 10:1, 1996.

A. Canavan, D. Graff, and G. Zipperlen. Callhome american english speech. *Linguistic Data Consortium*, 1997.

S. K. Card, T. P. Moran, and A. Newell. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1983. ISBN 0898592437.

R. Carter. Orders of reality: Cancode, communication, and culture. *ELT journal*, 52(1):43–56, 1998.

R. Carter and M. McCarthy. *Cambridge grammar of English: a comprehensive guide; spoken and written English grammar and usage*. Ernst Klett Sprachen, 2006.

T. L. Chartrand and J. A. Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893, 1999.

T. Chen and M. Kan. Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*, 47(2):299–335, 2013.

Y.-N. Chen, D. Hakkani-Tür, and X. He. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6045–6049. IEEE, 2016.

G. Churcher, E. S. Atwell, and C. Souter. *Dialogue management systems: a survey and overview*. University of Leeds, School of Computing Research Report 1997.06, 1997.

P. R. Cohen. If not turing's test, then what? *AI magazine*, 26(4):61, 2005.

K. M. Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4:515–534, 1981.

R. M. Cooper. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1):84–107, 1974.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

H. Cuayáhuitl, S. Keizer, and O. Lemon. Strategic dialogue management via deep reinforcement learning. *arXiv preprint arXiv:1511.08099*, 2015.

C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Association for Computational Linguistics, Workshop on Cognitive Modeling and Computational Linguistics*, 2011.

L. Daubigney, M. Geist, S. Chandramohan, and O. Pietquin. A comprehensive reinforcement learning framework for dialogue management optimization. *IEEE Journal of Selected Topics in Signal Processing*, 6(8):891–902, 2012.

M. Davies. Comparing the corpus of american soap operas, COCA, and the BNC, 2012a.

M. Davies. Corpus of american soap operas, 2012b.

I. de Kok, D. Heylen, and L. Morency. Speaker-adaptive multimodal prediction model for listener responses. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013.

L. Deng and X. Li. Machine learning paradigms for speech recognition: An overview. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(5):1060–1089, 2013.

L. Denoyer and P. Gallinari. The wikipedia xml corpus. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 12–19. Springer, 2007.

B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, and W. Cohen. Tweet2vec: Character-based distributed representations for social media. *arXiv preprint arXiv:1605.03481*, 2016.

A. Djalali, S. Lauer, and C. Potts. Corpus evidence for preference-driven interpretation. In *Logic, Language and Meaning*, pages 150–159. Springer, 2012.

J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*, 2015.

S. Dose. Flipping the script: A corpus of american television series (cats) for corpus-based language learning and teaching. *Corpus Linguistics and Variation in English: Focus on Non-native Englishes*, 2013.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pages 488–500. Springer, 2007.

W. Eckert, E. Levin, and R. Pieraccini. User modeling for spoken dialogue system evaluation. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 80–87, 1997.

L. El Asri, H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman. Frames: Acorpus for adding memory to goal-oriented dialogue systems. preprint on webpage at `http://www.maluuba.com/publications/`, 2017.

M. Elsner and E. Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *Association for Computational Linguistics (ACL)*, 2008.

M. Elsner and E. Charniak. Disentangling chat. *Computational Linguistics*, 36(3):389–409, 2010.

D. Erhan, Y. Bengio, A. Courville, Pierre-A. Manzagol, and P. Vincent. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11, 2010.

G. Di Fabbrizio, G. Tur, and D. Hakkani-Tr. Bootstrapping spoken dialog systems with data reuse. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2004.

M. Fatemi, L. E. Asri, H. Schulz, J. He, and K. Suleman. Policy networks with two-stage training for dialogue systems. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2016.

Maryam Fazel-Zarandi, Shang-Wen Li, Jin Cao, Jared Casale, Peter Henderson, David Whitney, and Alborz Geramifard. Learning robust dialog policies in noisy environments. *Neural Information Processing Systems (NIPS), Conversational AI Workshop*, 2017.

D. Fisher, M. Smith, and H. T Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 59b–59b, 2006.

P. Forchini. Spontaneity reloaded: American face-to-face and movie conversation compared. In *Corpus Linguistics*, 2009.

P. Forchini. *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Peter Lang, 2012.

G. Forgues, J. Pineau, J. Larchevêque, and R. Tremblay. Bootstrapping dialog systems with word embeddings. In *Workshop on Modern Machine Learning and Natural Language Processing, Neural Information Processing Systems (NIPS)*, 2014.

E. N. Forsyth and C. H. Martell. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC).*, pages 19–26, 2007.

M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 445–450, 2015.

M. Gašić, F. Jurčíček, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*. ACL, 2010.

M. Gašić, F. Jurčíček, B. Thomson, K. Yu, and S. Young. On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 312–317. IEEE, 2011.

M. Gašić, M. Henderson, B. Thomson, P. Tsiakoulis, and S. Young. Policy optimisation of pomdp-based dialogue systems without state space compression. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 31–36. IEEE, 2012.

M. Gasic, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young. On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371, 2013.

M. Gašić, N. Mrkšić, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young. Dialogue manager domain adaptation using gaussian process reinforcement learning. *Computer Speech & Language*, 2016.

A. Genevay and R. Laroche. Transfer learning for user adaptation in spoken dialogue systems. In *International Conference on Autonomous Agents & Multiagent Systems*, pages 975–983. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

K. Georgila, J. Henderson, and O. Lemon. User simulation for spoken dialogue systems: learning and evaluation. In *INTERSPEECH*, 2006.

K. Georgila, M. Wolters, J. D. Moore, and R. H. Logie. The MATCH corpus: A corpus of older and younger users interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44(3):221–261, 2010.

J. Gibson and A. D. Pick. Perception of another person's looking behavior. *The American journal of psychology*, 76(3): 386–394, 1963.

J. J Godfrey, E. C Holliman, and J McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, 1992.

I. Goodfellow, A. Courville, and Y. Bengio. Deep learning. Book in preparation for MIT Press, 2015. URL http://goodfeli.github.io/dlbook/.

C. Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic Press, 1981.

A. L. Gorin, G. Riccardi, and J. H. Wright. How may I help you? *Speech Communication*, 23(1):113–127, 1997.

A. Graves. Sequence transduction with recurrent neural networks. In *International Conference on Machine Learning (ICML), Representation Learning Workshop*, 2012.

S. Greenbaum. *Comparing English worldwide: The international corpus of English*. Clarendon Press, 1996.

S. Greenbaum and G Nelson. The international corpus of english (ICE) project. *World Englishes*, 15(1):3–15, 1996.

C. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535, 2015.

I. Gurevych and M. Strube. Semantic similarity applied to spoken dialogue summarization. In *International Conference on Computational Linguistics (COLING)*, 2004.

V. Haslerud and A. Stenström. The bergen corpus of london teenager language (COLT). *Spoken English on Computer. Transcription, Mark-up and Application. London: Longman*, pages 235–242, 1995.

P. A. Heeman and J. F. Allen. The TRAINS 93 Dialogues. Technical report, DTIC Document, 1995.

C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The ATIS spoken language systems pilot corpus. In *DARPA Speech and Natural Language Workshop*, pages 96–101, 1990.

M. Henderson, B. Thomson, and S. Young. Deep neural network approach for the dialog state tracking challenge. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2013.

M. Henderson, B. Thomson, and J. Williams. Dialog state tracking challenge 2 & 3, 2014a.

M. Henderson, B. Thomson, and J. Williams. The second dialog state tracking challenge. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2014b.

M. Henderson, B. Thomson, and S. Young. Word-based dialog state tracking with recurrent neural networks. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2014c.

P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. Ethical challenges in data-driven dialogue systems. *arXiv preprint arXiv:1711.09050*, 2017.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.a N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

T. Hiraoka, G. Neubig, K. Yoshino, T. Toda, and S. Nakamura. Active learning for example-based dialog systems. In *Proc Intl Workshop on Spoken Dialog Systems, Saariselka, Finland*, 2016.

H. Hung and G. Chittaranjan. The IDIAP wolf corpus: exploring group behaviour in a competitive role-playing game. In *International Conference on Multimedia*, pages 879–882, 2010.

J. L. Hutchens and M. D. Alder. Introducing MegaHAL. In *Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, 1998.

J. L. Part I. Shalyminov X. Xu Y. Yu O. Duek V. Rieser O. Lemon I. Papaioannou, A. C. Curry. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. In *Alexa Prize Proceedings*, 2017.

A. Jonsson and N. Dahlback. Talking to a computer is not like talking to your best friend. In *Scandinavian Conference on Artificial Intelligence*, 1988.

S. Jung, C. Lee, K. Kim, M. Jeong, and G. G. Lee. Data-driven user simulation for automated evaluation of spoken dialog systems. *Computer Speech & Language*, 23(4):479–509, 2009.

D. Jurafsky and J. H. Martin. *Speech and language processing, 2nd Edition*. Prentice Hall, 2008.

F. Jurcıcek, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu, and S. Young. Real user evaluation of spoken dialogue systems using amazon mechanical turk. In *INTERSPEECH*, volume 11, 2011.

R. Kadlec, M. Schmid, and J. Kleindienst. Improved deep learning baselines for ubuntu corpus dialogs. *Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*, 2015.

S. Kim, L. F. DHaro, R. E. Banchs, J. Williams, and M. Henderson. Dialog state tracking challenge 4, 2015.

S. Kim, L. F. DHaro, R. E. Banchs, J. D. Williams, M. Henderson, and K. Yoshino. The fifth dialog state tracking challenge. In *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

V. Konovalov, O. Melamud, R. Artstein, and I. Dagan. Collecting better training data using biased agent policies in negotiation dialogues. In *WOCHAT: Workshop on Chatbots and Conversational Agent Technologies*, 2016.

J. A Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. *Neural Information Processing Systems (NIPS)*, 2015.

M. Kytö and T. Walker. *Guide to A corpus of English dialogues 1560-1760*. Acta Universitatis Upsaliensis, 2006.

I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *Association for Computational Linguistics and International Conference on Computational Linguistics*, volume 1, pages 704–710. ACL, 1998.

C.-J. Lee, S.-K. Jung, K.-D. Kim, D.-H. Lee, and G. G. Lee. Recent approaches to dialog management for spoken dialog systems. *Journal of Computing Science and Engineering*, 4(1):1–22, 2010.

G. Leech. 100 million words of english: the british national corpus (BNC). *Language Research*, 28(1):1–13, 1992.

O. Lemon. Adaptive natural language generation in dialogue using reinforcement learning. In *Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, volume 8, pages 149–156, 2008.

E. Levin and R. Pieraccini. A stochastic model of computer-human interaction for learning dialogue strategies. In *Eurospeech*, volume 97, pages 1883–1886, 1997.

E. Levin, R. Pieraccini, and W. Eckert. Learning dialogue strategies within the markov decision process framework. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 72–79. IEEE, 1997.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

J. Li, M. Galley, C. Brockett, J. Gao, and Bill D. A persona-based neural conversation model. In *Association for Computational Linguistics*, pages 994–1003, 2016.

Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

G. Lin and M. Walker. All the world's a stage: Learning character models from film. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2011a.

G. I. Lin and M. A. Walker. All the world's a stage: Learning character models from film. In *AIIDE*, 2011b.

B. Liu and I. Lane. Multi-domain adversarial learning for slot filling in spoken language understanding. *Neural Information Processing Systems (NIPS), Conversational AI Workshop*, 2017.

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132, 2016.

C. Lord and M. Haith. The perception of eye contact. *Attention, Perception, & Psychophysics*, 16(3):413–416, 1974.

R. Lowe, N. Pow, I. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2015a.

R. Lowe, N. Pow, I. V. Serban, L. Charlin, and J. Pineau. Incorporating unstructured textual knowledge sources into neural dialogue systems. *Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*, 2015b.

R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. On the evaluation of dialogue systems with next utterance classification. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2016.

J. M. Lucas, F. Fernndez, J. Salazar, J. Ferreiros, and R. San Segundo. Managing speaker identity and user profiles in a spoken dialogue system. In *Procesamiento del Lenguaje Natural*, number 43 in 1, pages 77–84, 2009.

B. MacWhinney and C. Snow. The child language data exchange system. *Journal of child language*, 12(02):271–295, 1985.

F. Mairesse and S. Young. Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, 2014.

F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young. Phrase-based statistical language generation using graphical models and active learning. In *the Association for Computational Linguistics*, pages 1552–1561. ACL, 2010.

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

M. McCarthy. *Spoken language and applied linguistics*. Ernst Klett Sprachen, 1998.

S. McGlashan, N. Fraser, N. Gilbert, E. Bilange, P. Heisterkamp, and N. Youd. Dialogue management for telephone information systems. In *Conference on Applied natural Language Processing*, pages 245–246. ACL, 1992.

G. McKeown, M. F Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084, 2010.

G. Mesnil, X. He, L. Deng, and Y. Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775, 2013.

T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*, pages 3111–3119, 2013.

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

S. Miller, D. Stallard, R. Bobrow, and R. Schwartz. A fully statistical approach to natural language interfaces. In *Association for Computational Linguistics*, pages 55–61, 1996.

K. Mo, S. Li, Y. Zhang, J. Li, and Q. Yang. Personalizing a dialogue system with transfer learning. *arXiv preprint arXiv:1610.02891*, 2016.

S. Mohan and J. Laird. Learning goal-oriented hierarchical tasks from situated interactive instruction. In *AAAI*, 2014.

T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

L Nio, S. Sakti, G. Neubig, T. Toda, and S. Nakamura. Conversation dialog corpora from television and movie scripts. In *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–4, 2014.

E. Nöth, A. Horndasch, F. Gallwitz, and J. Haas. Experiences with commercial telephone-based dialogue systems. *it–Information Technology (vormals it+ ti)*, 46(6/2004):315–321, 2004.

C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28, 2013.

A. H. Oh and A. I. Rudnicky. Stochastic language generation for spoken dialogue systems. In *ANLP/NAACL Workshop on Conversational systems*, volume 3, pages 27–32. ACL, 2000.

T. Paek. Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment. In *Proc. Dialog-on-Dialog Workshop, INTERSPEECH*, 2006.

K. Papineni, S. Roukos, T Ward, and W Zhu. BLEU: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, 2002.

G. Parent and M. Eskenazi. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 312–317. IEEE, 2010.

A. N. Pargellis, H-K. J. Kuo, and C. Lee. An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, 42(3-4):329–351, 2004. doi: 10.1016/j.specom.2003.10.003.

R. Passonneau and E. Sachar. Loqui human-human dialogue corpus (transcriptions and annotations), 2014.

B. Peng, X. Li, L. Li, J. Gao, A. Celikyilmaz, S. Lee, and K.-F. Wong. Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2221–2230, 2017.

D. Perez-Marin and I. Pascual-Nieto. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, 2011.

S. Petrik. *Wizard of Oz Experiments on Speech Dialogue Systems*. PhD thesis, Technischen Universitat Graz, 2004.

R. Pieraccini, D. Suendermann, K. Dayanidhi, and J. Liscombe. Are we there yet? research in commercial spoken dialog systems. In *Text, Speech and Dialogue*, pages 3–13, 2009.

O. Pietquin and H. Hastie. A survey on metrics for the evaluation of user simulations. *The knowledge engineering review*, 28(01):59–73, 2013.

O. Pietquin, M. Geist, S. Chandramohan, and H. Frezza-Buet. Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(3):7, 2011.

B. Piot, M. Geist, and O. Pietquin. Imitation learning applied to embodied conversational agents. In *4th Workshop on Machine Learning for Interactive Systems (MLIS 2015)*, volume 43, 2015.

C. Potts. Goal-driven answers in the cards dialogue corpus. In *West Coast Conference on Formal Linguistics*, pages 1–20, 2012.

A. Ratnaparkhi. Trainable approaches to surface natural language generation and their application to conversational dialog systems. *Computer Speech & Language*, 16(3):435–455, 2002.

A. Raux, B. Langner, D. Bohus, A. W. Black, and M. Eskenazi. Lets go public! Taking a spoken dialog system to the real world. In *INTERSPEECH*, 2005.

N. Reithinger and M. Klesen. Dialogue act classification using language models. In *EuroSpeech*, 1997.

H. Ren, W. Xu, Y. Zhang, and Y. Yan. Dialog state tracking using conditional random fields. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2013.

S. Renals, T. Hain, and H. Bourlard. Recognition and understanding of meetings the AMI and AMIDA projects. In *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007.

R. Reppen and N. Ide. The american national corpus overall goals and the first release. *Journal of English Linguistics*, 32(2):105–113, 2004.

J. Rickel and W. L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied artificial intelligence*, 13(4-5):343–382, 1999.

V. Rieser and O. Lemon. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Empirical methods in natural language generation*, pages 105–120. Springer, 2010.

Verena Rieser and Oliver Lemon. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media, 2011.

A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, 2010.

A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

S. Rosenthal and K. McKeown. I couldnt agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2015.

S. Rosset and S. Petel. The ritel corpus-an annotated human-machine open-domain question answering spoken dialog corpus. In *The International Conference on Language Resources and Evaluation (LREC)*, 2006.

A. Roy, C. Guinaudeau, H. Bredin, and C. Barras. TVD: a reproducible and multiply aligned tv series dataset. In *The International Conference on Language Resources and Evaluation (LREC)*, volume 2, 2014.

J. Ruppenhofer, M. Ellsworth, M. R.L. Petruck, C. R. Johnson, and J. Scheffczyk. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, 2006. Distributed with the FrameNet data.

J. Schatzmann and S. Young. The hidden agenda user simulation model. *IEEE transactions on audio, speech, and language processing*, 17(4):733–747, 2009.

J. Schatzmann, K. Georgila, and S. Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2005.

J. Schatzmann, B. Thomson, K. Weilhammer, . Ye, and S. Young. 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*, pages 149–152, 2007.

K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *International Conference on Human Language Technology Research*, pages 12–19. Morgan Kaufmann Publishers Inc., 2002.

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

N. Schrading, C. O. Alm, R. Ptucha, and C. M. Homan. An analysis of domestic a.se discourse on reddit. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

K. K. Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005. Paper AAI3179808.

I. V. Serban and J. Pineau. Text-based speaker identification for multi-participant open-domain dialogue systems. *Neural Information Processing Systems Workshop on Machine Learning for Spoken Language Understanding*, 2015.

I. V. Serban, R. Lowe, L. Charlin, and J. Pineau. Generative deep neural networks for dialogue: A short review. In *Neural Information Processing Systems (NIPS), Let's Discuss: Learning Methods for Dialogue Workshop*, 2016.

I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Networks. In *AAAI*, 2016. In press.

I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI Conference*, 2017a.

I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, et al. A Deep Reinforcement Learning Chatbot. *arXiv preprint arXiv:1709.02349*, 2017b.

I. V Serban, C. Sankar, S. Zhang, Z. Lin, S. Subramanian, T. Kim, S. Chandar, N. R. Ke, et al. The octopus approach to the alexa competition: A deep ensemble-based socialbot. In *Alexa Prize Proceedings*, 2017c.

I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI Conference*, 2017d.

S. Shaikh, T. Strzalkowski, G. A. Broadwell, J. Stromer-Galley, S. M. Taylor, and N. Webb. Mpc: A multi-party chat corpus for modeling social phenomena in discourse. In *The International Conference on Language Resources and Evaluation (LREC)*, 2010.

L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

C. Shaoul and C. Westbury. A Usenet Corpus (2005-2009), 2009.

S. Sharma, J. He, K. Suleman, H. Schulz, and P. Bachman. Natural language generation in dialogue using lexicalized and delexicalized data. *arXiv preprint arXiv:1606.03632*, 2016.

B. A. Shawar and E. Atwell. Different measurements metrics to evaluate a chatbot system. In *Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 89–96, 2007a.

B. A. Shawar and Eric Atwell. Chatbots: are they really useful? In *LDV Forum*, volume 22, pages 29–49, 2007b.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI meeting recorder dialog act (mrda) corpus. Technical report, DTIC Document, 2004.

A. Simpson and N. M Eraser. Black box and glass box evaluation of the SUNDIAL system. In *European Conference on Speech Communication and Technology*, 1993.

S. Singh, D. Litman, M. Kearns, and M. Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, pages 105–133, 2002.

S. P. Singh, M. J. Kearns, D. J. Litman, and M. A. Walker. Reinforcement learning for spoken dialogue systems. In *Neural Information Processing Systems*, 1999.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, 2015.

A. Stenström, G. Andersen, and I. K. Hasund. *Trends in teenage talk: Corpus compilation, analysis and findings*, volume 8. J. Benjamins, 2002.

A. Stent, R. Prasad, and M. Walker. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Association for Computational Linguistics*, page 79. ACL, 2004.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

P.-H. Su, Y.-B. Wang, T.-H. Yu, and L.-S. Lee. A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8213–8217. IEEE, 2013.

P.-H. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T.-H. Wen, and S. Young. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *INTERSPEECH*, 2015.

P.-H. Su, M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*, 2016.

X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt, and M. Pantic. A multimodal database for mimicry analysis. In *Affective Computing and Intelligent Interaction*, pages 367–376. Springer, 2011.

J. Svartvik. *The London-Lund corpus of spoken English: Description and research*. Number 82 in 1. Lund University Press, 1990.

B. Thomson and S. Young. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588, 2010.

J. Tiedemann. Parallel data, tools and interfaces in opus. In *The International Conference on Language Resources and Evaluation (LREC)*, 2012.

D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2*, pages 766–773. ACM, 2002.

G. Tur and L. Deng. *Intent Determination and Spoken Utterance Classification*, chapter 4, pages 81–104. Wiley, January 2011.

A. M. Turing. Computing machinery and intelligence. *Mind*, pages 433–460, 1950.

D. C Uthus and D. W Aha. The ubuntu chat corpus for multiparticipant chat analysis. In *AAAI Spring Symposium: Analyzing Microtext*, 2013.

J. Vandeventer, A. J. Aubrey, P. L. Rosin, and D. Marshall. 4d cardiff conversation database (4D CCDb): A 4D database of natural, dyadic conversations. In *Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, 2015.

D. Vandyke, P.-H. Su, M. Gasic, N. Mrksic, T.-H. Wen, and S. Young. Multi-domain dialogue success classifiers for policy training. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 763–770. IEEE, 2015.

O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Paradise: A framework for evaluating spoken dialogue agents. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 271–280, 1997.

M. A. Walker, O. C. Rambow, and M. Rogati. Training a sentence planner for spoken dialogue using boosting. *Computer Speech & Language*, 16(3):409–433, 2002.

M. A. Walker, R. Grant, J. Sawyer, G. I. Lin, N. Wardrip-Fruin, and M. Buell. Perceived or not perceived: Film character models for expressive nlg. In *ICIDS*, pages 109–121, 2011.

M. A Walker, G. I. Lin, and J. Sawyer. An annotated corpus of film dialogue for learning and characterizing character style. In *The International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1378, 2012a.

M. A Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *The International Conference on Language Resources and Evaluation (LREC)*, pages 812–817, 2012b.

R. S. Wallace. The anatomy of alice. *Parsing the Turing Test*, pages 181–210, 2009.

Ye-Yi Wang, Li Deng, and Alex Acero. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5): 16–31, 2005.

Z. Wang and O. Lemon. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2013.

S. Webb. A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, 15(4):497–519, 2010.

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

T. Wen, M. Gašic, D. Kim, N. Mrkšic, P. Su, D. Vandyke, and S. Young. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2015a.

T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015b.

T.-H. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, and S. Young. Multi-domain neural network language generation for spoken dialogue systems. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2016)*, 2016.

T.-H. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young. A network-based end-to-end trainable task-oriented dialogue system. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2017.

J. Weston. Dialog-based language learning. *arXiv preprint arXiv:1604.06045*, 2016.

J. Williams, A. Raux, D. Ramachandran, and A. Black. The dialog state tracking challenge. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2013.

J. D. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

J. D. Williams and G. Zweig. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*, 2016.

M. Wolska, Q. B. Vo, D. Tsovaltzi, I. Kruijff-Korbayová, E. Karagjosova, H. Horacek, A. Fiedler, and C. Benzmüller. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *The International Conference on Language Resources and Evaluation (LREC)*, 2004.

B. Wrede and E. Shriberg. Relationship between dialogue acts and hot spots in meetings. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 180–185. IEEE, 2003.

X. Yang, Y.-N. Chen, D. Hakkani-Tür, P. Crook, X. Li, J. Gao, and L. Deng. End-to-end joint learning of natural language understanding and dialogue manager. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5690–5694. IEEE, 2017.

Y. Yang, W. Yih, and C. Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2013–2018, 2015.

Z. Yang, B. Li, Y. Zhu, I. King, G. Levow, and H. Meng. Collection of user judgments on spoken dialog system with crowdsourcing. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 277–282, 2010.

S. Young, M. Gasic, B. Thomson, and J. D. Williams. POMDP-based statistical spoken dialog systems: A review. *IEEE*, 101(5):1160–1179, 2013.

S. J. Young. Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769), 2000.

J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2016)*, 2016.