A Survey of Available Corpora For Building Data-Driven Dialogue Systems

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

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

Laurent Charlin {LCHARLIN} AT CS DOT MCGILL DOT CA
Department of Computer Science, McGill University
3480 University st, Montréal, QC H3A 0E9, Canada

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

Editor: N/A

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 and how they can be used to learn diverse dialogue strategies. We also describe other potential uses of these datasets, such as methods for transfer learning between datasets and the use of external knowledge, and discuss appropriate choice of evaluation metrics for the learning objective.

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 for dialogue such as dialogue policy learning (Young et al., 2013), dialogue state tracking (Williams et al., 2013; Kim et al., 2015; Henderson et al., 2013) and natural language generation (Wen et al., 2015).

We hypothesize that, in general, much of the recent progress is due to the availability of large public datasets, increased computing power, and development of new machine learning models, in particular neural network architectures. To facilitate further research on building data-driven dialogue systems, in this paper we therefore present a broad survey of available dialogue corpora together with an overview of corpus-based approaches for building dialogue systems.

Corpus-based learning is not the only approach to training dialogue systems. Researchers have also proposed trainng dialogue systems online through live interaction with humans, and off-line using user simulator models and reinforcement learning methods (Young et al., 2013; Gasic et al., 2013; Mohan and Laird, 2014; Pietquin and Hastie, 2013; Jung et al., 2009; Georgila et al., 2006). However, these approaches are beyond the scope of this survey.

In the following section, we discuss general characteristics of dialogue systems which are relevant for training data-driven dialogue systems. We discuss the purpose and goal of the dialogue system, the individual system components which are relevant for data-driven approaches, whether the system must be capable of long-term interactions, and whether the system communicates using written or spoken language. We also discuss the benefits and drawbacks of training data-driven dialogue systems based on different dialogue corpora, such as corpora containing dialogues between two humans or between a human and a machine, corpora constructed from natural or from unnatural dialogues, and corpora constructed from fictive dialogues. In Section 3 we present our survey over dialogue corpora according to the categories discussed in Section 2. In particular, we categorize the corpora based on whether dialogues are between humans or between a human and a machine, and whether the dialogues are in written or spoken language. 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 4 we discuss in detail the most popular data-driven approaches to training dialogue systems. We discuss various data pre-processing issues, such as spelling correction and speaker segmentation, discriminative model architectures, such as dialogue act classification, and response generation models. In Section 5 we conclude the survey by discussing issues related to corpus size and transfer learning between corpora, the incorporation of external knowledge into the dialogue system, data-driven learning for contextualization and personalization, and automatic evaluation metrics. Finally, we briefly discuss the future of data-driven dialogue systems.

2. Characteristics of Data-Driven Dialogue Systems

This section offers a broad characterization of data-driven dialogue systems, which is used throughout our presentation of datasets and learning architectures.

2.1 An Overview of Dialogue Systems

The standard architecture for dialogue systems, shown in Figure 1, incorporates a Speech Recognizer, Language Interpreter, State Tracker, Response Generator, Natural Language Generator, and Speech Synthesizer. In the case of text-based (written) dialogues, the Speech Recognizer and Speech Synthesizer can be left-out. While some of the literature on dialogue systems identifies only the State Tracker and Response Selection components as belonging inside the dialogue manager (Young, 2000), throughout this paper we adopt a broader view where the language understanding and generation are incorporated within the dialogue system. This leaves space for the development and analysis of end-to-end dialogue systems, which are starting to gain attention (Serban et al., 2015).

While there are substantial opportunities to improve each of the components in Figure 1 through data driven approaches, within this survey we focus primarily on datasets suitable to enhance the four components inside the Dialogue System box. It is worth noting that the Natural Language Interpreter and Generator are core NLP problems with applications well beyond dialogue systems.

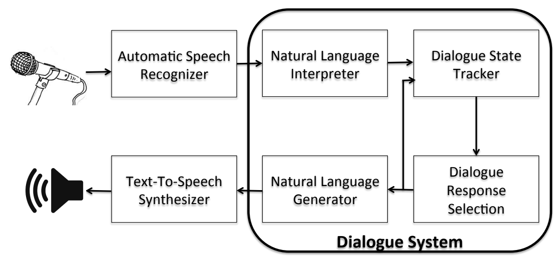

Figure 1: Dialogue System Diagram

2.2 Tasks and objectives

Dialogue systems have been built for a wide range of purposes. A useful distinction can be made between developing goal driven dialogue systems, such as technical support services, and non-goal driven dialogue systems, such as language learning tools or computer game characters. 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.

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 which managed to convincingly mimic a Rogerian psychotherapist by persistently either 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. Later work, for example by Hutchens and Alder (1998), started to develop data-driven systems (Shawar and Atwell, 2007b). Under the assumption that the coverage of different topics and general fluency is of primary importance, these systems were trained on a mixture of data sources ranging from both real and fictive dialogues to arbitrary texts. In comparison to the rule-based systems, these systems have had limited success (Perez-Marin and Pascual-Nieto, 2011; Shawar and Atwell, 2007b). Recent neural network architectures trained on large-scale corpora have shown promising results (Shang et al., 2015; Sordoni et al., 2015b; Vinyals and Le, 2015; Serban et al., 2015). However, they require having a sufficiently large dataset - in the hundreds of millions or even billions of words - in order to achieve these results.

Initial work on goal driven dialogue systems primarily used rule-based systems, similar to the non-goal driven systems referenced above, with the distinction that machine learning techniques have been heavily used to classify the intention (or need) of the user, as well as to bridge the gap between text and speech (Gorin et al., 1997). Research in this area started to take off during the mid 90s, 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 the non-goal driven systems, industry here 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 centres. 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 features (Young et al., 2013). In particular, the datasets are usually constrained to a very small task.

2.3 Discriminative Learning of Dialogue System Components

Modern dialogue systems consist of several components, as illustrated in Figure 1. Several of these components can be learned through discriminative models, which aim to predict labels or annotations relevant to other parts of the dialogue system.

An example of a discriminative model is a user intent classification model, which is trained to predict the intent of a user in a dialogue conditioned on the utterances of that user. In this case, the intent is called the *label*, *target* or *output*, and the conditioned utterances are called the *conditioning variables* or *inputs*. Discriminative models can be similarly applied in all parts of the dialogue system, including speech recognition, natural language understanding, state tracking and response selection.

Discriminative models fall into the machine learning paradigm of supervised learning. When the labels of interest are discrete, the models are called *classification* models, which is the most common case, and when the labels of interest are continuous they 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).

Discriminative models have allowed goal driven dialogue systems to make significant progress (Williams et al., 2013). 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 a dialogue state tracker). With proper annotations, discriminative models can be evaluated automatically and accurately.

2.4 Deterministic Vs. Generative Response Models

Dialogue systems with data-driven response generation components can be divided into two categories: those that select deterministically from a fixed set of possible responses, and those that attempt to generate responses by keeping a posterior over possible utterances. Those in the first category essentially skip the Natural Language Generator in Figure 1. In that case, the model maps the output of the dialogue tracker or natural language understanding modules together with the dialogue history (e.g. previous tracker outputs and previous system actions) and external knowledge (e.g. a database, which can be queried by the system) to a response action:

$$f_\theta \ : \{\text{dialogue history},\ \text{tracker outputs}, \text{external knowledge}\} \rightarrow \ \text{action}\ a_t, \tag{1}$$

where a_t is the dialogue system response action at time t, and θ is the set of parameters which defines f. Rule-based systems, such as ELIZA, use a deterministic mapping and therefore naturally belong to the category of response selection models. In such systems, the mapping function, f_θ, is usually a discrete function composed by aggregating all the response generation rules. Information retrieval and ranking-based systems, such as the model proposed by Banchs and Li (2012), which search through a database of dialogues and pick responses with the most similar context also belong to this category. Although the mapping function, f_θ, is still discrete for such models, the dialogue history and tracker outputs are usually projected into an Euclidean space (e.g. using TF-IDF bag-of-words representations) and a desirable response region is found (e.g. a point in the Euclidian space). The optimal 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. 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. (2015b), 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 probability of the recurrent neural network model. Although both of its sub-components are based on probabilistic models, the final model does not construct a probability distribution over all possible responses.

In contrast, a generative response model explicitly computes a full posterior probability distribution over possible system actions at every turn:

$$P_\theta(\text{action } a_t \mid \text{dialogue history, tracker outputs, external knowledge}).$$

Reinforcement learning systems with stochastic policies, such as the NJFun system (Singh et al., 2002), belong to this category. These systems typically have a very small set of possible (hand-crafted) system states and actions, which makes it possible to apply established reinforcement learning algorithms to train them either online or offline. 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" . Unfortunately, this severely limits their application area.

Systems based on recurrent neural networks also belong to this category (Serban et al., 2015; Vinyals and Le, 2015; Wen et al., 2015). By breaking down eq. (2) into a product of probabilities over words, they are able to generate responses by sampling word-by-word from their probability distribution. Unlike the deterministic response models, these systems are also able to generate entirely novel responses (e.g. by sampling word-by-word). They are further able to generate highly probable responses, i.e. the response with the highest probability, by using a method known as beam-search (Graves, 2012). These systems project each word into an Euclidian space (known as a word embedding) (Bengio et al., 2003); they also project the dialogue history and external knowledge into an Euclidian space (Wen et al., 2015; Lowe et al., 2015b). Although still in their infancy, similar models have been applied for statistical machine translation the last few years with promising results (Auli et al., 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014).

When trained solely on text, these generative models can be viewed as unsupervised learning models, because they aim to reproduce data distributions. In other words, the models learn to assign a probability to every possible conversation. Furthermore, since they generate responses word by word, they must learn to simulate the behaviour of the agents in the training corpus.

2.5 Long-term Interactions

Several applications require dialogue systems that are able to handle multiple interactive turns. For example, an interactive system might require steps of clarification from the user before being able to offer pertinent information. Indeed, this is a common scenario: many dialogues between humans, as well as between humans and machines, yield significant ambiguity (Heisterkamp, 1993), which can usually be resolved over the course of the dialogue. This phase is sometimes referred to as the grounding process (Clark and Brennan, 1991).

To tackle such behaviors, it is crucial to have access to dialogue corpora with long interactions, which include clarifications and confirmations which are ubiquitous in human conversations. The need for such long-term interactions is confirmed by recent empirical results, which show that longer interactions help generate appropriate responses (Serban et al., 2015; Vinyals and Le, 2015).

Corpora with long interactions may also be important for incorporating personalization in dialogue, e.g. adopting the language of the dialogue system to fit that of the user. To generate personalized responses to the user, it is useful to have access to either long-term dialogues or to dialogues with returning users, such as a user who interacts with a dialogue system in a smartphone application over the course of several days or months.

2.6 Spoken Vs. Written Language

Dialogue systems differ in the modalities through which they interact. An important distinction is whether they interact through written language, spoken language or in a multi-modal setting (e.g. by taking both audio and video as input). From a natural language perspective, there are key differences between written and spoken language. Spoken language tends to be less formal, has lower information content and contains many more pronouns than written language (Carter and McCarthy, 2006; Biber and Finegan, 2001, 1986). In particular, the differences are even larger 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, first person pronouns, second person pronouns, questions, contradictions as well as that-clauses and if-clauses appear very frequently in face-to-face conversations. Forchini (2012) summarized these differences by stating that: *"... studies show that face-to-face conversation is interpersonal, situation-dependent, and has no narrative concern or as Biber and Finegan (1986) put it, is a highly interactive, situated and immediate text type..."* In the following sections, we will therefore distinguish between dialogue corpora in written and spoken language.

2.7 Human-Human Vs. Human-System Corpora

One of the most important distinctions between dialogue datasets resides in the nature of the conversing parties—notably, whether it involves interactions between two humans, or between a human and a computer[1]. The distinction must be made as current artificial dialogue systems are significantly constrained, and of course do not produce nearly the same distribution of possible responses. As stated by (Williams and Young, 2007):

1. Machine-machine dialogue corpora are not of interest to us, because dialogues in general must have at least have one human participating to sufficiently reflect natural human language and because user simulation models are outside the scope of this survey.

> (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).

Williams and Young (2007) have previously argued against building data-driven dialogue systems using human-human dialogues by noting that: *"... using humanhuman conversation data is not appropriate because it does not contain the same distribution of understanding errors, and because humanhuman turn-taking is much richer than humanmachine dialog."*. 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, and to goal driven dialogue systems, where it is often possible to learn an effective dialogue model using reinforcement learning techniques. Williams and Young (2007) also argue against learning from corpora generated between humans or from existing dialogue systems: *"while it would be possible to use a corpus collected from an existing spoken dialogue system, supervised learning would simply learn to approximate the policy used by that spoken dialogue system and an overall performance improvement would therefore be unlikely."* Thus, for goal driven spoken dialogue systems in particular, it would appear that the most effective strategy is learning online through interaction with real users.

Nevertheless, in this survey we assume that we do not have the possibility of learning a dialogue system through online interaction with real users, and instead focus on data-driven learning from human-human and human-machine dialogue corpora. In general, human-human dialogue corpora may be more suitable for learning open domain dialogue systems, because they reflect natural dialogue interactions. By natural dialogues, we refer to conversations which are unconstrained and unscripted, e.g. interlocutors are not instructed to carry out a particular task, to follow a series of instructions, or to act out a scripted dialogue, and the dialogue process is relatively unaffected by researchers, e.g. interlocutors are not interrupted by question prompts in the middle of a dialogue. As can be expected, such conversations include a significant amount of turn-taking, pauses and common grounding phenomena (Clark and Brennan, 1991). Additionally, they are more diverse, and open up the possibility for the model to learn to understand natural language. Nonetheless there are interesting 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 learning an effective dialogue model. The goal is to learn a policy that is eventually better than the original stochastic policy used to generate the corpus (in a process known as bootstrapping).

2.8 Natural Vs. Unnatural Corpora

The way in which a dialogue corpus was generated and collected can have a significant influence on the trained data-driven dialogue system. In the case of human-human dialogues, an ideal corpus should very closely resemble natural dialogues between humans. Arguably, this is the case when conversations between humans are recorded and transcribed. The humans in the dialogue represent the true population of users with whom the dialogue system is intended to interact, and they are unaware of the fact that they are being recorded. This is not always possible due to ethical considerations and resource constraints.

Researchers may be forced to inform the human interlocutors that they are being recorded. Researchers are also often constrained 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 this case, there is no guarantee that the interactions with the dialogue system in the corpus will reflect true interactions, since the hired humans may behave differently from the true user population. One aspect that may cause behavioural differences is the fact that the hired humans may not share the same intentions and motivations as the true user population (Young et al., 2013). The unnaturalness may be further increased by the hiring process as well as the platform through which they interact with the dialogue system. However, when executed carefully, experiments based on crowdsourcing platforms, such as Amazon Mechanical Turk, may be able to evaluate spoken dialogue systems efficiently (Jurcıcek et al., 2011). The level of unnaturalness may be reduced when when the dialogue system with which they interact is operated by a human. These experiments are known as *Wizard-of-Oz* experiments (Bohus and Rudnicky, 2008; Petrik, 2004). These experiments generate datasets which are closer in nature to the dialogues humans would have with the real dialogue system, but unfortunately they are also very expensive and time-consuming to carry out. Ultimately the impact of any unnaturalness in the dialogues depends on the task and context in which the dialogue system will be deployed.

2.9 Corpora from Fiction

Naturally, it is also possible to use artificial dialogue corpora for data-driven learning, such as corpora based on works of fiction, including 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 may make data-driven learning harder as these events cannot be observed. The same problem is also observed in certain other media, such as microblogging websites (e.g. Twitter and Weibo), where conversations may also depend on external unobserved events.

Nevertheless, recent studies have found that spoken language in movies resembles human spoken language (Forchini, 2009). Although movie language is explicitly written to be spoken and contains certain artificial elements, many of the linguistic and paralinguistic features are similar to natural spoken language, including dialogue acts such as turn-taking and reciprocality (e.g. returning a greeting when greeted). The artificial differences that exist may even be helpful for data-driven dialogue learning, because movie dialogues are more compact, follow a steady rhythm and contain less garbling and repetition, yet still present a clear event or message to the viewer (Dose, 2013; Forchini, 2009, 2012). Unlike dialogues extracted from *Wizard-of-Oz* human experiments, movie dialogues also span many different topics and occur in many different environments (Webb, 2010). They contain different actors with different intentions and relationships to one another, which could potentially allow a data-driven dialogue system to learn to personalize itself to different users by making use of different interaction patterns. In comparing movie dialogues to actual human-human dialogues, Forchini (2012) concludes that: *"(a) movie conversation takes place in the spoken medium and occurs with non-verbal paralinguistic features; (b) movie conversation [...] [takes] place in real time, if real time is perceived as an ongoing process. Although movies are pre-recorded and not impromptu events, the audience perceives that something is happening while watching the movie [...] (c) movie conversation usually takes place in a shared context; (d) movie conversation is interactive, continuous, and expressive of politeness, emotion, and attitude."*

3. 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 to build dialogue systems, and to very large corpora of interactions—which may or may not be strictly considered dialogue datasets—that could be leveraged in the near future to build more sophisticated data-driven dialogue models. 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, and later highlight some of the more promising examples, explaining how they could be used to further dialogue research.

The dialogue datasets analyzed in this paper 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.

3.1 Human-Machine Corpora

As we discussed in Subsection 2.7, 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. One of the most popular 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 the Dialogue State Tracker (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. The challenge was created in order to standardize the process of comparing different state tracking models, and to ease the burden of data collection for new research groups.

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. DSTC1 (Williams et al., 2013) has 15,000 conversations with an automated bus ride system, where the requests of users are static and straightforward. DSTC2 introduces changing user goals in a restaurant booking system, and has 3,000 dialogues (Henderson et al., 2014b). DSTC3 introduces a small amount of labelled data in the domain of tourist information. It has only 2,265 dialogues, but 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-type booking system. It is a medium-sized dataset of 15,000 spoken conversations regarding up-to-the-minute flight information, hotel information, and car rentals. This information was transcribed, along with the user's comments at the end of the interaction. As with the DSTC datasets, this corpus could be used to enhance travel-oriented dialogue systems, or indeed any system that retrieves structured information to the user.

Name	Type	Topics	Avg. # of turns	Total # of dialogues	Total # of words	Description
DSTC1 (Williams et al., 2013)	Spoken	Buse schedules	–	15,000	–	Bus ride information system
DSTC2 (Henderson et al., 2014b)	Spoken	Restaurants	–	3,000	–	Restaurant booking system
DSTC3 (Henderson et al., 2014a)	Spoken	Tourist information	–	2,265	–	Information for tourists
CMU Communicator Corpus (Bennett and Rudnicky, 2002)	Spoken	Travel	11.67	15,481	–	Travel planning and booking system
ATIS Pilot Corpus (Hemphill et al., 1990)	Travel	Travel	25.4	41	10.6K-12.2K*	Travel planning and booking system
Ritel Corpus (Rosset and Petel, 2006)	Spoken	Unrestricted/ Diverse Topics	–	582	60k	An annotated open-domain question answering spoken dialogue system
DIALOG Mathematical Proofs (Wolska et al., 2004)	Spoken	Mathematics	12	66	–	Humans interact with computer system to do mathematical theorem proving
MATCH Corpus (Georgila et al., 2010)	Spoken	Appointment Scheduling	–	447	–	A system for scheduling appointments. Includes dialogue act annotations
MeMo Corpus (Möller et al., 2008)	Spoken	Household	–	64	–	Instructing smart home system

Table 1: Human-machine dialogue datasets. Starred (*) numbers give lower and upper bound on the quantity based on the average number of words per utterance.

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.
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 (Kim et al., 2015)	Tourist	35	273k	21hrs	Tourist information exchange over Skype.

Table 2: 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`).

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	–	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	–	500k	55hrs	Spontaneous teenage talk recorded in 1993. Conversations were recorded secretly.
Longman Spoken American Corpus (Stern, 2005)	Casual Topics	–	5M	550hrs*	First dataset to capture American spoken English on a large scale.
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	–	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	–	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 3: 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)

Name	Topics	Total # of utterances	Total # of dialogues	Total # of scripts	Total # of words	Description
Movie-DiC (Banchs, 2012)	Movie dialogues	764k	132,229	753	6M	Movie scripts of American films.
Movie-Triples (Serban et al., 2015)	Movie dialogues	736k	245,296	614	13M	Triples of utterances which are filtered to come from X-Y-X triples.
American Film Scripts	Movie scripts	1M*	–	1,500	16M*	Movie scripts of American films.
Cornell Movie-Dialogue Corpus (Danescu-Niculescu-Mizil and Lee, 2011)	Movie dialogues	1.4	220,579	617	9M*	Short conversations from film scripts, annotated with character metadata.
Filtered Movie Script Corpus (Nio et al., 2014b)	Movie dialogues	173k	86,719	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*	–	22,000	100M	Transcripts of American soap operas.
TVD Corpus (Roy et al., 2014)	TV show scripts	60k*	–	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	—	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*	–	207,907	1B	Movie subtitles which are not speaker-aligned.

Table 4: Human-human scripted dialogue datasets. 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).

Name	Type	Topics	Avg. # of turns	Total # of dialogues	Total # of words	Description
NPS Chat Corpus (Forsyth and Martell, 2007)	Chat	Unrestricted	–	–	100M	Posts from age-specific online chat rooms.
Twitter Corpus (Ritter et al., 2010)	Microblog	Unrestricted	2	1,300,000	–	Tweets and replies extracted from Twitter
Twitter Triple Corpus (Sordoni et al., 2015b)	Microblog	Unrestricted	3	29,000,000	–	A-B-A triples extracted from Twitter
Sina Weibo Corpus (Shang et al., 2015)	Microblog	Unrestricted	2	4,435,959	100M	Posts and replies extracted from Sina Weibo
Sina Weibo Short-Text Conversation Corpus (Wang et al., 2013)	Microblog	Unrestricted	2	15,000	–	Posts, along with unlabeled replies from Sina Weibo
UseNet Corpus (Shaoul and Westbury, 2009)	Microblog	Unrestricted	–	–	7B	UseNet forum postings
NUS SMS Corpus (Chen and Kan, 2013)	SMS messages	Unrestricted	–	–	580,668*□	SMS messages collected between two users, with timing analysis.
Reddit	Forum	Unrestricted	–	–	–	Comment trees from Reddit
Reddit Domestic Abuse Corpus (Schrading et al., 2015)	Forum	Abuse help	17.53	21,133	19M-103M △	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'
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	930,000	100M	Dialogues extracted from Ubuntu chat stream on IRC
Ubuntu Chat Corpus (Uthus and Aha, 2013)	Chat	Ubuntu Operating System	–	–	2B*□	Chat stream scraped from IRC logs (no dialogues extracted)
Movie Dialog Dataset (Dodge et al., 2015)	Chat, QA & Recommendation	Movies	3.3	3.1M▼	–	For goal-driven dialogue systems. Includes movie metadata as knowledge triples.

Table 5: Human-human written dialogue datasets. Starred (*) quantities are computed using word counts based on spaces (i.e. a word must be a sequence of characters preceded and followed by a space), but for certain corpora, such as IRC and SMS corpora, proper English words are sometimes concatenated together due to slang usage. Triangle (△) indicates lower and upper bounds computed using average words per utterance estimated on a similar Reddit corpus Schrading (2015). Square (□) indicates estimates based only on English part of the corpus. Note that 2.1M dialogues from the Movie Dialog dataset (▼) are in the form of simulated QA pairs.

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

The **Ritel corpus** (Rosset and Petel, 2006) is a small dataset of 528 spoken questions and answers in a conversational format. The purpose of the project 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, such as "Who is currently presiding the Senate?", along with some conversations about arts and science-related topics.

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 collected 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.

A similar goal motivated the development of the **MeMo corpus** (Möller et al., 2008), which records the interactions of older and younger users with a dialogue system for controlling domestic devices. Both interactions and subsequent user questionnaires were collected in a *Wizard-of-Oz* experiment, involving 32 users in a laboratory setting. User keywords and subdialogues corresponding to task boundaries were also annotated.

3.2 Human-Human Spoken Corpora

Naturally, there is much more data available from conversations between humans as opposed to conversations between a human and a machine. Thus, we break down this category further, into spoken dialogues (this section) and written dialogues (Section 3.3). 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 2.6, spoken dialogues tend to be more colloquial, use shorter words and phrases and are generally less well-formed, as the user is speaking in a train-of-thought manner. Conversely, in written communication, users have the ability to reflect on what they are writing before they send a message. Written dialogues can also contain spelling errors or abbreviations, which are generally not transcribed in spoken dialogues.

3.2.1 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 and 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 over the phone from 500 speakers, along with word-by-word transcriptions. 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 of other 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. It 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) has 60 telephone conversations lasting 5-30 minutes each, between English speakers in North America without a Southern accent. It is annotated with speaker information such as sex, age, and education. The goal of the project was to support the development of language identification technologies; however, 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 has been used to identify trends in language evolution for teenagers (Stenström et al., 2002).

As most previously available large spoken datasets used British English, the **Longman Spoken American Corpus** (Stern, 2005) claims to be the first to document American spoken English on a substantial scale. It was intended to capture the colloquial nature of American English, and was used to create a "Dictionary of Contemporary English". It is roughly half of the size of the BNC. There is also a parallel dataset, collected by the same authors, entitled the Longman British Spoken Corpus, which is a subset of the BNC.

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).

Other work has 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, including 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, naturalistic—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, it is a small dataset of multi-participant chats, and it has not been disentangled into strict dialogue. It has been used for analysis of the dynamics of various corporate and academic meeting scenarios.

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. Their original dataset consisted of 30 five minute conversations, 7 of which are fully annotated with transcriptions and behavioural annotations such as speaker activity, facial expressions, head motions, and smiles. The content of the conversation is free chatting 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 in the late 1960s-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. It 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 dialogue; the latter refers to dialogue that is 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 have been normalised 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, made from 1948-1961 in order to document various existing English dialects. People aged 60 and over were recruited, as they are most likely to speak the traditional 'uncontaminated' dialects of their area. They were encouraged to talk about their memories, families, work, and the folklore of their countryside from a century ago.

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 contains transcripts, with occasional audio and video recordings, of data collected from children and adults who are 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 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, which 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.

3.2.2 Constrained Spoken Corpora

Next, we discuss domains in which conversations only occur about a particular topic, or intend to solve a specific task. Not only is the topic of the conversation specified beforehand, but participants are discouraged or disincentivize 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 dicussed in Subsection 2.8 this may also make the conversations less natural.

A well-known example of such a dataset is the **HCRC Map Task Corpus** (Anderson et al., 1991), which consists of unscripted, task-oriented dialogues which 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 on one of the participant's map to the map of another participant. The corpus is fairly small, with 64 dialogues containing approximately 150,000 words. It attempts to support the study of human speech, and controls for the familiarity of speakers, eye contact between speakers, matching between landmarks on the participants' maps, opportunities for contrastive stress, and phonological characteristics of landmark names.

Other corpora impose similar direction-giving goals between participants. 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 was to provide a testbed for natural lexical entrainment, and to be used as a resource for pedestrian navigation applications.

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 **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 of speech involving over 400 speakers, recorded between 1994-1998. The CPASE has two main components; the first is transcripts (0.9 million words) of press conferences from the White House, which contains almost exclusively question and answer sessions in addition to some policy statements by politicians. The second consists of transcripts (1.1 million words) of faculty meetings and committee meetings related to national tests, which involve statements, discussions and questions. The creation of the corpus was motivated by the desire to understand and model more formal uses of the English language.

There are many constrained corpora that provide multi-modal data for analyzing human behaviour during conversations. These often consists of auditory, visual, and written transcriptions of the dialogues. One such example is the **MAHNOB Mimicry Database** (Sun et al., 2011), which is a set of fully synchronised audio-visual recordings of natural dyadic (one-on-one) interactions. It contains 11 hours of recordings, split over 54 sessions conversations between 60 people, who are engaged either in a socio-political discussion or negotiating a tenancy agreement. 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.

Similarly, the **IDIAP Wolf Corpus** (Hung and Chittaranjan, 2010) is also an audio-visual corpus containing natural conversational data of volunteers who took part in an adversarial role-playing game called 'Werewolf'. Four groups of 8-12 people were recorded using headset microphones and synchronised 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 were recorded while holding conversations with an operator, who adopted various roles designed to evoke emotional reactions. These were also filmed and recorded with synchronous video and audio devices. Importantly, the operators' responses were stock phrases which 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.

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, the **DSTC 4** (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 from other datasets in the DSTC in order to improve state tracking performance.

3.2.3 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 2.9, these 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, these scripted dialogues are required to be dramatic, as they are generally sourced from movies or TV shows.

There exists 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, where consecutive utterances are not divided or labeled in any way. Of course, it is always preferable to have the speaker labels; however, there is significantly more unlabeled subtitle data available, and both sources of information can be leveraged in building 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[2], 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., 2015). There is also the **American Film Scripts Corpus**, which can be purchased online[3], and the **Cornell Movie-Dialogue Corpus** which can be downloaded free of charge.[4]

The majority of these datasets consist mostly of raw scripts, which are not guaranteed to portray conversations between only two people. The dataset collected by Nio et al. (2014b), which we refer to as the **Filtered Movie Script Corpus**, takes over 1 million utterance-response pairs from web-based script resources, and filter them to 86,000 such pairs. The filtering method limits the extracted utterances to X-Y-X triples, where X is spoken by the same actor and each of the utterance share some semantic similarity, which 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.

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 consist of single utterances.

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 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 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, this includes aligning an event description with all the utterances associated with that event. This could be used to develop algorithms for locating specific events from raw dialogue, such as ’person X tries to convince person Y’.

2. http://www.imsdb.com
3. http://alexanderstreet.com/products/american-film-scripts
4. http://www.mpi-sws.org/~cristian/Cornell_Movie-Dialogs_Corpus.html

Some work has been done in order to analyze character style from movie scripts. This is aided by a dataset collected in (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.[5] 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, which can help dialogue systems deal with out-of-domain (OOD) interactions.

3.3 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[6], where users can make posts and other users can make comments or replies to a post. In some cases, comments can be nested indefinitely, as users make replies to previous replies. Utterances 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[7] 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.

3.3.1 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 was annotated with part-of-speech and dialogue act information; the correctness of this was verified manually. It 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) containing 1.3 million post-reply pairs extracted from Twitter. It was originally constructed to aid in the production of unsupervised approaches to modeling dialogue acts. Even larger corpora such as the **Twitter Triples Corpus** (Sordoni et al., 2015b),

5. http://www.opensubtitles.org
6. http://www.reddit.com
7. http://www.twitter.com

which has 127 million context-message-response triples, and the **Sina Weibo Corpus** (Shang et al., 2015), which contains 4.5 million post-reply pairs, have been collected; however, they have not yet been made publicly available. The **Sina Weibo Short-Text Conversation Corpus** (Wang et al., 2013) is a smaller section of Sina Weibo data, with 46,345 posts and 1.5 million responses, used to develop an information retrieval method for response generation. Unfortunately, only 12,000 of these responses are labeled with the corresponding posts to which they replied.

The **Usenet Corpus** (Shaoul and Westbury, 2009) is a gigantic collection of public Usenet postings[8], containing over 7 billion words, from October 2005 to January 2011. Usenet is a distributed discussion system established in 1980, where participants can post articles to one of 47,860 'newsgroup' categories. It is seen as the precursor to many current Internet forums. The 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. Unfortunately, the corpus does not consist of dialogues, but rather single SMS messages. SMS messages are similar in style to Twitter, using many abbreviations and acronyms.

Currently, one of the most popular forum-based websites is Reddit[9], 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**.[10] 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. Although researchers have not yet investigated dialogue problems using this Reddit discussion corpus, the sheer size of the dataset renders it an interesting candidate for transfer learning. Additionally, researchers have used smaller collections of Reddit discussions for broad discourse classification. (Schrading et al., 2015).

Some more curated versions of the Reddit dataset have been collected. 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 stopwords, and semantic role labels are provided.

3.3.2 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 to achieve a pre-specified number of points. Therefore, the game is adversarial in nature,

8. http://www.usenet.net
9. http://www.reddit.com
10. https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

and can be used to analyze situations of strategic conversation where the agents have diverging motives.

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 were 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), with 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.[11] 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. Thus, this dataset consists of similarly technical conversations 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)[12] of 116k examples with accompanying movie and actor metadata in the form of knowledge triples; a recommendation dataset from MovieLens[13] 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

11. http://irclogs.ubuntu.com
12. http://en.omdb.org
13. http://movielens.org

dataset, the dialogues in the sub-datasets are simulated QA pairs, where each response corresponds to a list of entities from the knowledge base.

4. Learning from Dialogue Corpora

In this section, we review existing computational architectures suitable for learning dialogue strategies directly from data. The goal is not to provide full technical details on the methods available to achieve this, though we provide appropriate citations for the interested reader. Rather, we aim to illustrate concretely how the datasets described above can, and have, been used in different dialogue learning efforts.

4.1 Data Pre-processing

Before applying machine learning methods to a dialogue corpus, it is common practice to perform some form of pre-processing. The aim of pre-processing is to standardize a dataset with minimal loss of information. This can reduce data scarcity, and eventually make it easier for models to learn from the dataset. In natural language processing, it is commonly acknowledged that pre-processing can have a significant effect on the results of the natural language processing system—the same observation holds for dialogue. Although the specific procedure for pre-processing is task- and data-dependent, in this section we highlight a few common approaches, in order to give a general idea of where pre-processing can be effective for dialogue systems.

Pre-processing is often used to remove anomalies in the data. For text-based corpora, this can include removing acronyms, slang, misspellings and phonemicization (e.g. where words are written according to their pronunciation instead of their correct spelling). For some models, such as the generative dialogue models discussed later, tokenization (e.g. defining the smallest unit of input) is also critical. In datasets collected from mobile text, forum, microblog and chat-based settings, it is common to observe a significant number of acronyms, abbreviations, and phonemicizations that are specific to the topic and userbase (Clark, 2003). Although there is no widely accepted standard for handling such occurrences, many NLP systems incorporate some form of pre-processing to normalize these entries (Kaufmann and Kalita, 2010; Aw et al., 2006; Clark, 2003). For example, there are look-up tables, such as the IRC Beginner List[14], which can be used to translate the most common acronyms and slang into standard English. Another common strategy is to use stemming and lemmatization to replace many words with a single item (e.g. *walking* and *walker* both replaced by *walk*). Of course, depending on the task at hand and the corpus size, an option is also to leave the acronyms and phonemicized words as they are.

In our experience, almost all dialogue datasets contain some amount of spelling errors. By correcting these, we expect to reduce data sparsity. This can be done by using automatic spelling correctors. However, it is important to inspect their effectiveness. For example, for movie scripts, Serban et al. (2015) found that automatic spelling correctors introduced more spelling errors than they corrected, and a better strategy was to use Wikipedia's most commonly misspelled words[15] to lookup and replace potential spelling errors. Transcribed spoken language corpora often include many non-words in their transcriptions (e.g. uh, oh). Depending on whether or not these provide

14. http://www.ircbeginner.com/ircinfo/abbreviations.html
15. https://en.wikipedia.org/wiki/Commonly_misspelled_English_words

additional information to the dialogue system, researchers may also want to remove these words by using automatic spelling correctors.

4.2 Segmenting Speakers and Conversations

Some dialogue corpora, such as those based on movie subtitles, come without explicit speaker segmentation. However, it is often possible to estimate the speaker segmentation, which is useful to build a model of a given speaker—as compared to a model of the conversation as a whole. For text-based corpora, Serban and Pineau (2015) have recently proposed the use of recurrent neural networks to estimate turn-taking and speaker labels in movie scripts with promising results.

In the speech recognition literature, this is the subtask of speaker diarisation (Miro et al., 2012; Tranter et al., 2006). When the audio stream of the speech is available, the segmentation is quite accurate with classification error rates as low as 5%.

A strategy sometimes used for segmentation of spoken dialogues is based on labelling a small subset of the corpus, known as the gold corpus, and training a specific segmentation model based on this. The remaining corpus is then segmented iteratively according to the segmentation model, after which the gold corpus is expanded with the most confident segmentations and the segmentation model is retrained. This process is sometimes known as embedded training, and is widely used in other speech recognition tasks (Jurafsky and Martin, 2008). It appears to work well in practice, but has the disadvantage that the interpretation of the label can drift. Naturally, this approach can be applied to text dialogues as well in a straightforward manner.

In certain corpora, such as those based on chat channels or extracted from movie subtitles, many conversations occur in sequence. In some cases, there are no labels partitioning the beginning and end of separate conversations. Similarly, certain corpora with multiple speakers, such as corpora based on chat channels, contain several conversations occurring in parallel (e.g. simultaneously) but do not contain any segmentation separating these conversations. This makes it hard to learn a meaningful model from such conversations, because they do not represent consistent speakers or coherent semantic topics.

To leverage such data towards learning individual conversations, researchers have proposed methods to automatically estimate segmentations of conversations (Lowe et al., 2015a; Nio et al., 2014a). Former solutions were mostly based on hand-crafted rules and seemed to work well upon manual inspection. For chat forums, one solution involves thresholding the beginning and end of conversations based on time (e.g. delay of more than x minutes between utterances), and eliminating speakers from the conversation unless they are referred to explicitly by other speakers (Lowe et al., 2015a). More advanced techniques involve maximum-entropy classifiers, which leverage the content of the utterances in addition to the discourse structure and timing information (Elsner and Charniak, 2008). For movie scripts, researchers have proposed the use of simple information-retrieval similarity measures, such as cosine similarity, to identify conversations (Nio et al., 2014a). Based on the their performance on estimating turn-taking and speaker labels, recurrent neural networks also hold promise for segmenting conversations (Serban and Pineau, 2015).

4.3 Discriminative Model Architectures

As discussed in Subsection 2.3, discriminative models aim to predict certain labels or annotations manually associated with a portion of a dialogue. For example, a discriminative model might be trained to predict the intent of a person in a dialogue, or the topic, or a specific piece of information.

In the following subsections, we discuss research directions where discriminative models have been developed to solve dialogue-related tasks.[16] This is primarily meant to review and contrast the work from a data-driven learning perspective.

4.3.1 Dialogue Act Classification and Dialogue Topic Spotting

Here we consider the simple task known as dialogue act classification (or dialogue move recognition). In this task, the goal is to classify a user utterance, independent of the rest of the conversation, as one out of K dialogue acts: $P(A \mid U)$, where A is the discrete variable representing the dialogue act and U is the user's utterance. This falls under the general umbrella of text classification tasks, though its application is specific to dialogue. Like the dialogue state tracker model, a dialogue act classification model could be plugged into a dialogue system as an additional natural language understanding component.

Early approaches for this task focused on using n-gram models for classification (Reithinger and Klesen, 1997; Bird et al., 1995). For example, Reithinger et al. assumed that each dialogue act is generated by its own language model. They trained an n-gram language model on the utterances of each dialogue act, $P_\theta(U|A)$, and afterwards use Bayes' rule to assign the probability of a new dialogue act $P_\theta(A|U)$ to be proportional to the probability of generating the utterance under the language model $P_\theta(U|A)$.

However, a major problem with this approach is the lack of datasets with annotated dialogue acts. More recent work by Forgues et al. (2014) acknowledged this problem, and tried to overcome the data scarcity issue by leveraging word embeddings learned from other, larger text corpora. They created an utterance-level representation by combining the word embeddings of each word, for example, by summing the word embeddings or taking the maximum w.r.t. each dimension. These utterance-level representations, together with word counts, were then given as inputs to a linear classifier to classify the dialogue acts. Thus, Forgues et al. showed that by leveraging another, substantially larger, corpus they were able to improve performance on their original task.

This makes the work on dialogue act classification very appealing from a data-driven perspective. First, it seems that the accuracy can be improved by leveraging alternative data sources. Second, unlike the dialogue state tracking models, dialogue act classification models typically involve relatively little feature hand-crafting thus suggesting that data-driven approaches may be more powerful for these tasks.

4.3.2 Dialogue State Tracking

The core task of the DSTC (Williams et al., 2013) adds more complexity by focusing on tracking the state of a conversation. This is framed as a classification problem: for every time step t of the dialogue, the model is given the current input to the dialogue state tracker (including ASR and SLU outputs) together with external knowledge sources (e.g. bus timetables). The required output is a probability distribution over a set of N_t predefined hypotheses, in addition to the REST hypothesis (which represents the probability that none of the previous N_t hypotheses are correct). The goal is to match the distribution over hypotheses as closely as possible to the real annotated

16. It is important to note that although discriminative models have been favored to model supervised problems in the dialogue-system literature, in principle generative models ($P(X,Y)$) instead of discriminative models ($P(Y|X)$) could be used.

data. By providing an open dataset with accurate labels, it has been possible for researchers to perform rigourous comparative evaluations of different classification models for dialogue systems.

Models for the DSTC include both statistical approaches and hand-crafted systems. An example of the latter is the system proposed in Wang and Lemon (2013), which relies on having access to a marginal confidence score $P_t(u, s, v)$ for a user dialogue $u(s = v)$ with slot s and value v given by a subsystem at time t. The marginal confidence score gives a heuristic estimate of the probability of a slot taking a particular value. The model must then aggregate all these estimates and confidence scores to compute probabilities for each hypothesis.

In this model, the SLU component may for example give the marginal confidence score (*inform(data.day=today)=0.9*) in the bus scheduling DSTC, meaning that it believes with high confidence (0.9) that the user has requested information for the current day. This marginal confidence score is used to update the belief state of the system $b_t(s, v)$ at time t using a set of hand-crafted updates to the probability distribution over hypotheses. From a data-driven learning perspective, this approach does not make efficient use of the dataset, but instead relies heavily on the accuracy of the hand-crafted tracker outputs.

More sophisticated models for the DSTC take a dynamic Bayesian approach by modeling the latent dialogue state and observed tracker outputs in a directed graphical model (Thomson and Young, 2010). These models are sometimes called generative state tracking models, though they are still discriminative in nature as they only attempt to model the state of the dialogue and not the words and speech acts in each dialogue. For simplicity we drop the index i in the following equations. Similar to before, let x_t be the observed tracker outputs at time t. Let s_t be the dialogue state at time t, which represents the state of the world including, for example, the user actions (e.g. defined by slot-value pairs) and system actions (e.g. number of times a piece of information has been requested). For the DSTC, the state s_t must represent the true current slot-value pair at time t. Let r_t be the reward observed at time t, and let a_t be the action taken by the dialogue system at time t. This general framework, also known as a partially-observable Markov decision process (POMDP) then defines the graphical model:

$$P_\theta(x_t, s_t, r_t | a_t, s_{t-1}) = P_\theta(x_t | s_t, a_t) P_\theta(s_t | s_{t-1}, a_t) P_\theta(r_t | s_t, a_t), \tag{2}$$

where a_t is assumed to be a deterministic variable of the dialogue history. This variable is given in the DSTC, because it comes from the policy used to interact with the humans when gathering the datasets. This approach is attractive from a data-driven learning perspective, because it models the uncertainty (e.g. noise and ambiguity) inherent in all variables of interest. Thus, we might expect such a model to be more robust in real applications.

Now, since all variables are observed in this task, and since the goal is to determine s_t given the other variables, we are only interested in:

$$P_\theta(s_t | x_t, r_t, a_t) \propto P_\theta(x_t | s_t, a_t) P_\theta(s_t | s_{t-1}, a_t) P_\theta(r_t | s_t, a_t), \tag{3}$$

which can then be normalized appropriately since s_t is a discrete stochastic variable. However, due to the temporal dependency between s_t and s_{t-1}, the complexity of the model is similar to a hidden Markov model, and thus both learning and inference become intractable when the state, observation and action spaces are too large. Indeed, as noted by Young et al. (2013), the number of states, actions and observations can easily reach 10^{10} configurations in some dialogue systems. Thus, it is necessary to make simplifying assumptions on the distribution $P_\theta(s_t | x_t, r_t, a_t)$ and to approximate the learning and inference procedures (Young et al., 2013). With appropriate structural assumptions and approximations, these models perform well compared to baseline systems on the DSTC (Black et al., 2011).

Non-bayesian data-driven models have also been proposed. These models are sometimes called discriminative state tracking models, because they do not assume a generation process for the tracker outputs, x_t or for any other variables, but instead only condition on them. For example, Henderson et al. (2013) proposed to use a feed-forward neural network. At each time step t, they extracted a set of features and then concatenate a window of W feature vectors together. These are given as input to the neural network, which outputs the probability of each hypothesis from the set of hypotheses. By learning a discriminative model and using a window over the last time steps, they do not face the intractability issues of dynamic Bayesian networks. Instead, their system can be trained with gradient descent methods. This approach could eventually scale to large datasets, and is therefore very attractive for data-driven learning. However, unlike the dynamic Bayesian approaches, these models do not represent probability distributions over variables apart from the state of the dialogue. Without probability distributions, it is not clear how to define a confidence interval over the predictions. Thus the models might not provide adequate information to determine when to seek confirmation or clarification following unclear statements.

Researchers have also investigated the use of conditional random fields (CRFs) for state tracking (Ren et al., 2013). This class of models also falls under the umbrella of discriminative state tracking models; however, they are able to take into account temporal dependencies within dialogues by modeling a complete joint distribution over states:

$$P_\theta(S|X) \propto \prod_{c \in C} \prod_i f_i(\mathbf{s_c}, \mathbf{x_c}), \tag{4}$$

where C is the set of factors, i.e. sets of state and tracker variables across time, $\mathbf{s_c}$ is the set of states associated with factor c, $\mathbf{x_c}$ is the set of observations associated with factor c, and $\{f_i\}_i$ is a set of functions parametrized by parameters θ. There exist certain functions f_i, for which exact inference is tractable and learning the parameters θ is efficient (Koller and Friedman, 2009; Serban). For example, Ren et al. (2013) propose a set of factors which create a linear dependency structure between the dialogue states while conditioning on all the observed tracker outputs:

$$P_\theta(S|X) \propto \prod_t \prod_i f_i(\mathbf{s_{t-1}}, \mathbf{s_t}, \mathbf{s_{t+1}}, X). \tag{5}$$

This creates a dependency between all dialogue states, forcing them be coherent with each other. This should be contrasted to the feed-forward neural network approach, which does not enforce any sort of consistency between different predicted dialogue states. The CFR models can be trained with gradient descent to optimize the exact log-likelihood, but exact inference is typically intractable. Therefore, an approximate inference procedure, such as loopy belief propagation, is necessary to approximate the posterior distribution over states s_t.

In summary, there exist different approaches to building discriminative learning architectures for dialogue. While they are fairly straightforward to evaluate and often form a crucial component for real-world dialogue systems, by themselves they only offer a limited view of what we ultimately want to accomplish with dialogue models. They often require labeled data, which is often difficult to acquire on a large scale (except in the case of answer re-ranking) and require manual feature selection, which reduces their potential effectiveness. Since each model is trained independently of the other models and components with which it interacts in the complete dialogue system, one cannot give guarantees on the performance of the final dialogue system by evaluating the individual models alone. Thus, we desire models that are capable of producing probability distributions over all possible responses instead of over all annotated labels—in other words, models that can actually *generate* new responses by selecting the highest probability next utterance. This is the subject of the next section.

4.4 Response Generation Models

Both the response re-ranking approach and the generative response model approach have allowed for the use of large-scale unannotated dialogue corpora for training dialogue systems. We therefore close this section by discussing these classes of approaches

In general, approaches which aim to generate responses have the potential to learn semantically more powerful representations of dialogues compared to models trained for dialogue state tracking or dialogue act classification tasks: the concepts they are able to represent are limited only by the content of the dataset, unlike the dialogue state tracking or dialogue act classification models which are limited by the annotation scheme used (e.g. the set of possible slot-value pairs pre-specified for the DSTC).

4.4.1 Re-ranking Response Models

Researchers have recently turned their attention to the problem of building models that produce answers by re-ranking a set of candidate answers, and outputting the one with the highest rank or probability. While the task may seem artificial, the main advantage is that it allows the use of completely un-annotated datasets. Unlike dialogue state tracking, this task does not require datasets where experts have labeled every utterance and system response. This task only requires knowing the sequence of utterances, which can be extracted automatically from transcribed conversations.

Banchs and Li (2012) construct an information retrieval system based on movie scripts using the vector space model. Their system searches through a database of movie scripts to find a dialogue similar to the current dialogue with the user, and then emits the response from the closest dialogue in the database. Similarly, Ameixa et al. (2014) also use an information retrieval system, but based on movie subtitles instead of movie scripts. They show that their system gives sensible responses to questions, and that bootstrapping an existing dialogue system from movie subtitles improves answering out-of-domain questions. Both approaches assume that the responses given in the movie scripts and movie subtitle corpora are appropriate. Such information retrieval systems consist of a relatively small set of manually tuned parameters. For this reason, they do not require (annotated) labels and can therefore take advantage of raw data (in this case movie scripts and movie subtitles). However, these systems are effectively nearest-neighbor methods. They do not learn rich representations from dialogues which can be used, for example, to generalize to previously unseen situations. Furthermore, it is unclear how to transform such models into full dialogue agents. They are not robust and it is not clear how to maintain the dialogue state. Contrary to search engines, which present an entire page of results, the dialogue system is only allowed to give a single response to the user.

(Lowe et al., 2015a) also propose a re-ranking approach using the Ubuntu Dialogue Corpus. The authors propose an *affinity model* between a context c (e.g. five consecutive utterances in a conversation) and a potential reply r. Given a context-reply pair the model compares the output of a context-specific LSTM against that of a response-specific LSTM neural network and outputs whether or not the response is correct for the given context. The model maximizes the likelihood of a correct context-response pair:

$$\max_{\theta} \sum_{i} P_{\theta}(\text{true response} \mid c_i, r_i)^{I_{c_i}(r_i)} (1 - P_{\theta}(\text{true response} \mid c_i, r_i))^{1-I_{c_i}(r_i)} \tag{6}$$

where θ stands for the set of all model parameters and $I_{c_i}(\cdot)$ denotes a function that returns 1 when r_i is the correct response to c_i and 0 otherwise. Learning in the model uses stochastic gradient descent. As is typical with neural network architectures, this learning procedure scales to large datasets. Given a context, the trained model can be used to pick an appropriate answer from a set of potential answers. This model assumes that the responses given in the corpus are appropriate (i.e., this model does not generate novel responses). However, unlike the above information retrieval systems, this model is not provided with a similarity metric as in the vector space model, but instead must learn the semantic relevance of a response to a context. This approach is more attractive from a data-driven learning perspective because it uses the dataset more efficiently and avoids costly hand tuning of parameters.

4.4.2 Full Generative Response Models

Generative dialogue response strategies are designed to automatically produce utterances by composing text (see Section 2.4). A straightforward way to define the set of dialogue system actions is by considering them as sequences of words which form utterances. Sordoni et al. (2015b) and Serban et al. (2015) both use this approach. They assume that both the user and the system utterances can be represented by the same generative distribution:

$$P_{\theta}(u_1, \ldots, u_T) = \prod_{t=1}^{T} P_{\theta}(u_t \mid u_{<t}) \tag{7}$$

$$= \prod_{t=1}^{T} \prod_{n=1}^{N} P_{\theta}(w_{t,n} \mid w_{t,<n}, u_{<t}), \tag{8}$$

where the dialogue consists of T utterances $u_1, \ldots, u_T$ and $w_{t,n}$ is the n^{th} token in utterance t. The variable $u_{<t}$ indicates the sequence of utterances which preceed u_t and similarly for $w_{t,<n}$. Further, the probability of the first utterance is defined as $P(u_1|u_{<1}) = P(u_1)$, and the first word of each utterance only conditions on the previous utterance, i.e. $w_{t,<1}$ is "null". Tokens can be words, as well as speech and dialogue acts. The set of tokens depends on the particular application domain, but in general the set must be able to represent all desirable system actions. In particular, the set must contain an end-of-utterance token to allow the model to express turn-taking. This approach is similar to language modeling. For differentiable models, training is based on maximum log-likelihood using stochastic gradient descent methods. As discussed in Subsection 2.4, these models project words and dialogue histories onto an Euclidian space. Furthermore, when trained on text only, they can be thought of as unsupervised machine learning models.

Sordoni et al. (2015b) use the above approach to generate responses for posts on Twitter. Specifically, $P_{\theta}(u_m \mid u_{<m})$ is given by a recurrent neural network which generates a response word-by-word based on Eq. (8). The model learns its parameters using stochastic gradient descent on a corpus of Twitter messages. The authors then combine their generative model with a machine translation

system and demonstrate that the hybrid system outperforms a state-of-the-art machine translation system (Ritter et al., 2011).

Serban et al. (2015) extend the above model to generate responses for movie subtitles and movie scripts. Specifically, Serban et al. (2015) adapt a hierarchical recurrent neural network (Sordoni et al., 2015a), which they argue is able to represent the common ground between the dialogue interlocutors. They also propose to add speech and dialogue acts to the vocabulary of the model to make the interaction with the system more natural. However, since the model is used in a standalone manner, i.e., without combining it with a machine translation system, the majority of the generated responses are highly generic (e.g. *I'm sorry* or *I don't know*). The authors conclude that this is a limitation of all neural network-based generative models for dialogue (e.g., (Serban et al., 2015; Sordoni et al., 2015b; Vinyals and Le, 2015)). The problem appears to lie in the distribution of words in the dialogue utterances, which primarily consist of pronouns, punctuation tokens and a few common verbs but rarely nouns, verbs and adjectives. When trained on a such a skewed distribution, the models do not learn to represent the semantic content of dialogues very well. This issue is exacerbated by the fact that dialogue is inherently ambiguous and multi-modal, which makes it more likely for the model to fall back on a generic response. As a workaround, Li et al. (2015) increase response diversity by changing the objective function at generation time to also maximize the mutual information between the context, i.e. the previous utterances, and the response utterance. However, it is not clear what impact this artificial diversity has on the effectiveness or naturalness of the dialogue system. It is possible that the issue may require larger corpora to learn semantic representations of dialogue, more context (e.g. longer conversations, user profiles and task-specific corpora) and multi-modal interfaces to reduce uncertainty. Further research is needed to resolve this question.

Wen et al. (2015) train a neural network to generate natural language responses for a closed-dialogue domain. They use Amazon Mechanical Turk[17] to collect a dataset of dialogue acts and utterance pairs. They then train recurrent neural networks to generate a single utterance as in Eq. (8), but condition on the specified dialogue act:

$$P_\theta(U|A) = \prod_n P_\theta(w_n \mid w_{<n}, A), \tag{9}$$

where A is the dialogue act represented by a discrete variable, U is the generated utterance given A and w_n is the n^{th} word in the utterance. Based on a hybrid approach combining different recurrent neural networks for answer generation and convolutional neural networks for re-ranking answers, they are able to generate diverse utterances representing the dialogue acts in their datasets.

Similar to the models which re-rank answers, generative models may be used as complete dialogue systems or as response generation components of other dialogue systems. However, unlike the models which re-rank answers, the word-by-word generative models can generate entirely new utterances never seen before in the training set. Further, in certain models such as those cited above, response generation scales irrespective of dataset size.

4.5 User Simulation Models

In the absence of large datasets, some researchers have turned to building user simulation models (sometimes referred to as 'user models') to train dialogue strategies. User simulation models aim

17. http://www.mturk.com

to produce natural, varied and consistent interactions from a fixed corpus, as stated by Pietquin and Hastie (2013, p. 2): "An efficient user simulation should not only reproduce the statistical distribution of dialogue acts measured in the data but should also reproduce complete dialogue structures." As such, they model the conditional probability of the user utterances given previous user and system utterances:

$$P_\theta(u_t^{\text{user}} | u_{<t}^{\text{user}}, u_{<t}^{\text{system}}), \tag{10}$$

where θ are the model parameters, u_t^{user} and u_t^{system} are the user utterance (or action) and the system utterance (or action) respectively at time t. Similarly, $u_{<t}^{\text{user}}$ and $u_{<t}^{\text{system}}$ indicate the sequence of user and system utterances that precede u_t^{user} and u_t^{system}, respectively.

There are two main differences between user simulation models and the generative response models discussed in Subsection 4.4.2. First, user simulation models never model the distribution over system utterances, but instead only model the conditional distribution over user utterances given previous user and system utterances. Second, user simulation models usually model dialogue acts as opposed to word tokens. Since a single dialogue act may represent many different utterances, the models generalize well across paraphrases. However, training such user simulation models requires access to a dialogue corpus with annotated dialogue acts, and limits their application to training dialogue systems which work on the same set of dialogue acts. For spoken dialogue systems, user simulation models are usually combined with a model over speech recognition errors based on the automatic speech recognition system but, for simplicity, we omit this aspect in our analysis.

Researchers initially experimented with n-gram-based user simulation models (Eckert et al., 1997; Georgila et al., 2006), which are defined as:

$$P_\theta(u_t^{\text{user}} | u_{t-1}^{\text{system}}, u_{t-2}^{\text{user}}, \ldots, u_{t-n-1}^{\text{system}}) = \theta_{u_t^{\text{user}}, u_{t-1}^{\text{system}}, u_{t-2}^{\text{user}}, \ldots, u_{t-n-1}^{\text{system}}}, \tag{11}$$

where n is an even integer, and θ is an n-dimensional tensor (table) which satisfies:

$$\sum_{u_t^{\text{user}}} \theta_{u_t^{\text{user}}, u_{t-1}^{\text{system}}, u_{t-2}^{\text{user}}, \ldots, u_{t-n-1}^{\text{system}}} = 1. \tag{12}$$

These models are trained either to maximize the log-likelihood of the observations by setting $\theta_{u_t^{\text{user}}, u_{t-1}^{\text{system}}, u_{t-2}^{\text{user}}, \ldots, u_{t-n-1}^{\text{system}}}$ equal to (a constant times) the number of occurrences of each corresponding n-gram , or on a related objective function which encourages smoothness and therefore reduces data sparsity for larger n's (Goodman, 2001). Even with smoothing, n has to be kept small and these models are therefore unable to maintain the history and goals of the user over several utterances (Schatzmann et al., 2005). Consequently, the goal of the user changes over time, which has a detrimental effect on the performance of the dialogue system trained using the user simulator.

Several solutions have been proposed to solve the problem of maintaining the history of the dialogue. Pietquin (2004) propose to condition the n-gram model on the user's goal:

$$P_\theta(u_t^{\text{user}} | u_{t-1}^{\text{system}}, u_{t-2}^{\text{user}}, \ldots, u_{t-n-1}^{\text{system}}, g), \tag{13}$$

where g is the goal of the user defined as a set of slot-value pairs. Unfortunately, not only must the goal lie within a set of hand-crafted slot-value pairs, but its distribution when simulating must also be defined by experts. Using a more data-driven approach, Georgila et al. (2006) propose to condition the n-gram model on additional features:

$$P_\theta(u_t^{\text{user}} | u_{t-1}^{\text{system}}, u_{t-2}^{\text{user}}, \ldots, u_{t-n-1}^{\text{system}}, f(u_{<t}^{\text{user}}, u_{<t}^{\text{system}})), \tag{14}$$

where $f(u_{<t}^{\text{user}}, u_{<t}^{\text{system}})$ is a function mapping all previous user and system utterances to a low-dimensional vector that summarizes the previous interactions between the user and the system (e.g. slot-value pairs that the user has provided the system up to time t). Now, θ can be learned using maximum log-likelihood with stochastic gradient descent.

More sophisticated probabilistic models have been proposed based on directed graphical models, such as hidden Markov models and input-output hidden Markov models (Cuayáhuitl et al., 2005), and undirected graphical models, such as conditional random fields based on linear chains (Jung et al., 2009). Inspired by Pietquin (2005), Pietquin (2007) and Rossignol et al. (2011) propose the following directed graphical model:

$$P_\theta(u_t^{\text{user}} | u_{<t}^{\text{user}}, u_{<t}^{\text{system}}) = \sum_{g_t, k_t} P_\theta(u_t^{\text{user}} | g_t, k_t, u_{<t}^{\text{user}}, u_{<t}^{\text{system}}) P_\theta(g_t | k_t) P_\theta(k_t | k_{<t}, u_{<t}^{\text{user}}, u_{<t}^{\text{system}}) \tag{15}$$

where g_t is a discrete random variable representing the user's goal at time t (e.g. a set of slot-value pairs), and k_t is another discrete random variable representing the user's knowledge at time t (e.g. a set of slot-value pairs). This model allows the user to change goals during the dialogue, which would be the case, for example, if the user is notified by the dialogue system that the original goal cannot be accomplished. The dependency on previous user and system utterances for u_t^{user} and k_t may be limited to a small number of previous turns as well as a set of hand-crafted features computed on these utterances. For example, the conditional probability:

$$P_\theta(u_t^{\text{user}} | g_t, k_t, u_{<t}^{\text{user}}, u_{<t}^{\text{system}}), \tag{16}$$

may be approximated by an n-gram model with additional features as in Georgila et al. (2006). Generating user utterances can be done in a straightforward manner by using ancestral sampling: first, sample k_t given $k_{<t}$ and the previous user and system utterances; then, sample g_t given k_t; and finally, sample u_t^{user} given g_t, k_t and the previous user and system utterances. The model can be trained using maximum log-likelihood. If all variables are observed, i.e. g_t and k_t have been given by human annotators, then the maximum-likelihood parameters can be found similarly to n-gram models by counting the co-occurrences of variables. If some variables are missing, they can be estimated using the expectation-maximization (EM) algorithm, since the dependencies form a linear chain. Rossignol et al. (2011) also propose to regularize the model by assuming a Dirichlet distribution prior over the parameters, which is straightforward to combine with the EM algorithm.

User simulation models are particularly useful in the development of dialogue systems based on reinforcement learning methods (Singh et al., 2002; Schatzmann et al., 2006; Pietquin and Dutoit, 2006; Frampton and Lemon, 2009; Jurčíček et al., 2012; Png and Pineau, 2011; Young et al., 2013). Furthermore, many user simulation models, such as those trainable with stochastic gradient descent or co-occurrence statistics, are able to scale to large corpora. In the light of the increasing availability of large dialogue corpora, there are ample opportunities for building novel user simulation models, which aim to better represent real user behavior, and in turn for training dialogue systems, which aim to solve more general and more difficult tasks. Despite their similarities, research on user simulation models and full generative models has progressed independently of each other so far. Therefore, it also seems likely that there is fruitful work to be done in transferring and merging ideas between these two areas.

5. Discussion

We conclude by discussing a number of general issues related to the development and evaluation of data-driven dialogue systems. We also discuss alternative sources of information, user personalization and automatic evaluation methods.

5.1 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 may be critical for building a data-driven dialogue system which performs well (Lowe et al., 2015a; Serban et al., 2015).

In particular, it is generally believed that the corpus size should scale with the number of topics covered. Indeed, 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 number of topics (e.g. misunderstanding the interpretation of a key word), the dataset needs to grow with the number of topics.

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 ambiguity and missing modalities (e.g. as in OpenSubtitles or Movie-Scriptolog datasets). If the corpus is sufficiently large, then the resolved ambiguities and missing modalities may, for example, be approximated using latent stochastic variables.

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 (Forgues et al., 2014). Indeed, in several branches of machine learning, and in particular in deep learning, the use of related datasets in pre-training the model is an effective method of scaling up to complex environments (Erhan et al., 2010).

On a related note, to build open-domain dialogue systems, it is arguably necessary to move beyond domain-specific datasets. Instead, like humans, the 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.

5.3 Building goal driven dialogue systems

Tables 1–5 list the topics of available datasets. Several of the human-human datasets are denoted as having Casual or Unrestricted Topics. This is in contrast to most of the human-machine datasets, which focus on specific topics. It can be useful to make a distinction, as goal driven dialogue systems typically have a well-defined measure of performance that is explicitly related to task completion. This information can be incorporated directly in the learning objective when building a data-driven dialogue system. In contrast, when building non-goal-driven dialogue systems, designed more to engage or entertain, the objective often is to simply carry out the conversation, but there is no notion of task completion. In some cases, the line between these two types of datasets blurs, for example in the case of conversations occurring between players of an online game (Afantenos et al., 2012): the outcome of the game is determined by their play in the game environment, not by their conversation, yet some elements of the conversation may have a direct impact on a player's performance in the game. In general, we highlight the distinction simply to emphasize that when a task completion metric is available, it is useful to incorporate it in the learning objective.

5.4 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 task-oriented setting (Raux et al., 2005; Nöth et al., 2004). Even chat-oriented 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, where there is not enough dialogue training data to reliably learn a response that is appropriate for each input utterance, or in domains that evolve quickly over time.

5.4.1 Structured External Knowledge

In traditional task-oriented *pipeline* 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 showtimes 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 Interpreter 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.

5.4.2 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 possible ways that these unstructured knowledge sources have or could be used in conjunction with a data-driven dialogue system, but we expect to see new results in this area in coming years.

5.4.3 Incorporating longer memories

Recently, significant progress has been made towards incorporating a form of external memory into various neural-network architectures for sequence modeling. Models such as Memory Networks (Weston et al., 2015; Sukhbaatar et al., 2015) and Neural Turing Machines (NTM) (Graves et al., 2014) store some part of their input in a memory, which is then reasoned over in order to perform a variety of sequence to sequence tasks. These vary from simple problems, such as sequence copying, to more complex problems, such as question answering and machine translation. Although none of these models are explicitly designed to address dialogue problems, the extension by Kumar et al. (Kumar et al., 2015) to Dynamic Memory Networks specifically differentiates between episodic and semantic memory. In this case, the episodic memory is the same as the memory used in the traditional Memory Networks paper which is extracted from the input, while the semantic memory refers to knowledge sources that are fixed for all inputs. The model is shown to work for a variety of NLP tasks, and it is not difficult to envision an application to dialogue utterance generation where the semantic memory is the desired external knowledge source.

5.5 Personalized dialogue agents

When conversing, humans often adapt to their interlocutor to facilitate understanding, and thus improve conversational efficiency and satisfaction. Thus, attaining human-level performance with dialogue agents may well require personalization, i.e. models that are aware and capable of adapting to their intelocutor. Such capabilities could increase the effectiveness and naturalness of generated dialogues (Lucas et al., 2009). We see personalization of dialogue systems as an important task, which so far has been mostly untouched. We do note that 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). 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.

5.6 Evaluation metrics

One of the most challenging aspects of constructing dialogue systems is their evaluation. While the end goal is usually to deploy the dialogue system in an application and receive real human feedback, getting to this stage is time consuming and expensive, and its usually 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 would be very inconvenient to re-run experiments for every parameter setting in a grid search. Although crowdsourcing platforms, such as Amazon Mechanical Turk, might prove to be a solution to these problems (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 by which to calculate a score for each model, and only once the best model has been chosen with reasonable confidence should human evaluators be involved.

The problem appears to be even harder for non-goal driven dialogue systems, where 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 each other's true identity. The goal of the judges 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 schemas (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). However, even when human evaluators are available, it is difficult to determine 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 utterance on vague notions such as '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. 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.

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

5.6.1 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, but more 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, automatic 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, simulated 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 the goal completion rate. Significant effort has been made to render the simulated data as realistic as possible, by properly 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.

5.6.2 Automatic Evaluation Metrics for Non-Goal Driven Dialogue Systems

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

Word Similarity Metrics. One approach is to borrow evaluation metrics from other NLP tasks such as machine translation, which uses BLEU (Papineni et al., 2002) or 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, given some context (Sordoni et al., 2015b). While BLEU scores have been shown to correlate with human judgements for machine translation (Papineni et al., 2002), their effectiveness for automatically assessing dialogue response generation is unclear. There are several issues: given the context of a conversation, there are a large number of possible responses that can be expected which 'fit' into the dialogue. Thus, the response generated by a dialogue system could be entirely reasonable, yet have no words in common with the actual next utterance. In this case, the BLEU score would be very low, but may 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., 2015b). Although the METEOR metric takes into account synonyms and morphological variants of the words in the candidate reply, it still suffers from the aforementioned problems. In a sense, these measurements only satisfy one direction of Schatzmann's criteria: while high BLEU and METEOR scores imply that the model is generating human-like output, the converse is not true. Such metrics will only accurately reflect the performance of the dialogue system if given a large number of candidate responses to each given context.

Next Utterance Classification. Alternatively, one can narrow down 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., 2015a). This *next utterance classification* (NUC) task is derived from the recall and precision metrics for information-retrieval based approaches. There are several attractive properties of this metric: it is easily interpretable, and one can adjust the difficulty by changing the number of false responses. However, there are drawbacks; in particular, 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 this metric by minimizing the cross-entropy between context-response pairs (Lowe et al., 2015a; Kadlec et al., 2015), the metric could also be used to evaluate a probabilistic generative model trained to output 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., 2015). 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, in which stop-words, punctuation, and end-of-utterance tokens are removed before evaluating, to focus on the semantic content of the phrase (Serban et al., 2015). 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 both produce human-like output and to reproduce most types of conversations in the corpus.

Response Diversity. As discussed in Subsection 4.4.2, recent non-goal driven dialogue systems based on neural networks have had problems generating diverse responses (Serban et al., 2015). (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

There is strong evidence that over the next few years, dialogue research will quickly move towards large-scale data-driven model approaches, in particular in the form of end-to-end trainable systems as is the case for other language-related applications such as speech recognition, machine translation and information retrieval. This paper provides an extensive survey of currently available datasets suitable for research, development, and evaluation of data-driven dialogue systems.

In addition to presenting the datasets, we provide an overview of existing work on learning dialogue system components and system strategies using various classes of approaches, from dialogue act classification and state tracking models to full generative response and user simulation models. While in many domains data scarcity poses important challenges, several potential extensions, such as transfer learning and incorporation of external knowledge, may provide scalable solutions.

References

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

Stergos 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.

Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. M., and K. Yamada. Translating with scarce resources. In *Association for the Advancement of Artificial Intelligence (AAAI/IAAI)*, 2000.

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

D. Ameixa, Luisa Coheur, P. Fialho, and P. Quaresma. Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. In *Intelligent Virtual Agents*, pages 13–21, 2014.

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.

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.

M. Auli, M. Galley, C. Quirk, and G. Zweig. Joint language and translation modeling with recurrent neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

A. Aw, M. Zhang, J. Xiao, and J. Su. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING*, pages 33–40, 2006.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*, 2015.

R. E. Banchs. Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the 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 *Proceedings of the ACL 2012 System Demonstrations*, 2012.

S. Banerjee and A. Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

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.

S. Bird, S. Browning, R. Moore, and M. Russell. Dialogue move recognition using topic spotting techniques. In *Spoken Dialogue Systems-Theories and Applications*, 1995.

A. W. Black, S. Burger, A. Conkie, H. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, et al. Spoken dialog challenge 2010: Comparison of live and control test results. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2011.

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

S. E. Brennan, K. S. Schuhmann, and K. M. Batres. Entrainment on the move and in the lab: The walking around corpus. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 2013.

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

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.

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.

Tanya 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.

K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

A. Clark. Pre-processing very noisy text. In *Proc. of Workshop on Shallow Processing of Large Corpora*, pages 12–22, 2003.

H. H. Clark and S. E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13:127–149, 1991.

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.

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. Renals, O. Lemon, and H. Shimodaira. Human-computer dialogue simulation using hidden markov models. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 290–295, 2005.

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

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. 2007.

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. 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.

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

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.

D. Fisher, M. Smith, and H. T Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual 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. 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. Frampton and O. Lemon. Recent research advances in reinforcement learning in spoken dialogue systems. *The Knowledge Engineering Review*, 24(04):375–408, 2009.

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. pages 445–450, 2015.

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.

K. Georgila, J. Henderson, and O. Lemon. User simulation for spoken dialogue systems: learning and evaluation. In *Proceedings of 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. 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/`.

J. T. Goodman. A bit of progress in language modeling extended version. *Machine Learning and Applied Statistics Group Microsoft Research. Technical Report, MSR-TR-2001-72*, 2001.

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 *Proceedings of the 29th International Conference on Machine Learning (ICML), Representation Learning Workshop*, 2012.

A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

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. volume abs/1503.03535, 2015.

I. Gurevych and M. Strube. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, 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. Heisterkamp. Ambiguity and uncertainty in spoken dialogue. In *In Proceedings Of The Third European Conference on Speech Communication and Technology (EUROSPEECH 1993)*, 1993.

C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The atis spoken language systems pilot corpus. In *Proceedings of the 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 *SIGDIAL*, 2014b.

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.

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

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

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. 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 *Proceedings of INTERSPEECH*, volume 11, 2011.

F. Jurčíček, B. Thomson, and S. Young. Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech & Language*, 26(3):168–192, 2012.

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.

M. Kaufmann and J. Kalita. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*, 2010.

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

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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.

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

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

G. Lin and M. Walker. All the world's a stage: Learning character models from film, 2011a.

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

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.

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, pages 77–84, 2009.

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

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.

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.

T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *11th Proceedings of 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 *Advances 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.

X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):356–370, 2012.

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

S. Möller, F. Gödde, and M. Wolters. A corpus analysis of spoken smart-home interactions with older users. 2008.

L. Nio, S. Sakti, G. Neubig, T. Toda, M. Adriani, and Sa. Nakamura. Developing non-goal dialog system based on examples of drama television. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 355–361. 2014a.

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, 2014b.

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.

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 *Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL)*, 2002.

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.

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, Citeseer, 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. *A framework for unsupervised learning of dialogue strategies*. Presses Université Catholique de Louvain, 2004.

O. Pietquin. A probabilistic description of man-machine spoken communication. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 410–413, 2005.

O. Pietquin. Learning to ground in spoken dialogue systems. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–165, 2007.

O Pietquin and T. Dutoit. A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):589–599, 2006.

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

S. Png and J. Pineau. Bayesian reinforcement learning for pomdp-based dialogue systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2156–2159, 2011.

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 *Proceedings of INTERSPEECH*. Citeseer, 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.

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

A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, 2011.

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.

S. Rossignol, O. Pietquin, and M. Ianotto. Training a bn-based user model for dialogue simulation with missing data. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 598–604, 2011.

A. Roy, C. Guinaudeau, H. Bredin, and C. Barras. Tvd: a reproducible and multiply aligned tv series dataset. The International Conference on Language Resources and Evaluation (LREC), 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, K. Georgila, and S. Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2005.

J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(02):97–126, 2006.

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.

J. N. Schrading. Analyzing domestic abuse using natural language processing on social media data, 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 *Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

K. K. Schuler. VerbNet: A broad-coverage, comprehensive verb lexicon. 2005. Ph.D. thesis. Paper AAI3179808.

I. V. Serban. Maximum likelihood learning and inference in conditional random fields. Master's thesis, University of Copenhagen, Denmark.

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, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Networks. In *AAAI Conference on Artificial Intelligence*, 2015. In press.

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.

B. A. Shawar and E. Atwell. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the 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.

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, Y. Bengio, H. Vahabi, C. Lioma, J. G. Simonsen, and J. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, 2015a.

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)*, 2015b.

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

K. Stern. The longman spoken american corpus: providing an in-depth analysis of everyday english. *AGE*, 18(24):20, 2005.

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.

S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *Neural Information Processing Systems (NIPS)*, 2015.

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. 2011.

I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, 2014.

J. Svartvik. *The London-Lund corpus of spoken English: Description and research*. Number 82. 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.

S. E. Tranter, D. Reynolds, et al. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.

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 *Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, 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 *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280, 1997.

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.

H. Wang, Z. Lu, Hang Li, and E. Chen. A dataset for research on short-text conversations. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

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)*, 2015.

J. Weston, S. Chopra, and A. Bordes. Memory networks. In *International Conference on Learning Representations (ICLR)*, 2015.

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.

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.

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. *Proceedings of the 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.