Big Data for Conversational Interfaces: Current Opportunities and Prospects

David Griol, Jose M. Molina and Zoraida Callejas

Abstract As conversational technologies develop, we demand more from them. For instance, we want our conversational assistants to be able to solve our queries in multiple domains, to manage information from different usually unstructured sources, to be able to perform a variety of tasks, and understand open conversational language. However, developing the resources necessary to develop systems with such capabilities demands much time and effort, as for each domain, task or language, data must be collected, annotated following an schema that is usually not portable, the models must be trained over the annotated data, and their accuracy must be evaluated. In recent years, there has been a growing interest in investigating alternatives to manual effort that allow exploiting automatically the huge amount of resources available in the web. In this chapter we describe the main initiatives to extract, process and contextualize information from these rich and heterogeneous sources for the various tasks involved in dialog systems, including speech processing, natural language understanding and dialog management.

Keywords Conversational interfaces · Big Data · Spoken interaction · Automatic speech recognition · Dialog management · Speech synthesis

D. Griol (✉) · J.M. Molina
Department of Computer Science, Carlos III University of Madrid, Avda, de la Universidad, 30, 28911 Leganés, Spain
e-mail: david.griol@uc3m.es

J.M. Molina
e-mail: josemanuel.molina@uc3m.es

Z. Callejas
Department of Languages and Computer Systems, University of Granada, CITIC-UGR, C/ Pdta. Daniel Saucedo Aranda S/n, 18071 Granada, Spain
e-mail: zoraida@ugr.es

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_6

1 Introduction

Speech and natural language technologies allow users to communicate in a flexible and efficient way, also enabling the access to applications when traditional input and output interfaces cannot be used (e.g. in-car applications, access for disabled persons, etc.). Also speech-based interfaces work seamlessly with small devices (e.g., smarthphones and tablets PCs) and allow users to easily invoke local applications or access remote information. For this reason, spoken dialog systems [22, 48, 59] are becoming a strong alternative to traditional graphical interfaces which might not be appropriate for all users and/or applications.

These systems are computer programs that receive speech as input and generate as output synthesized speech, engaging the user in a dialog that aims to be similar to that between humans [48, 59]. Thus, these interfaces make technologies more usable, as they ease interaction [23], allow integration in different environments [22], and make technologies more accessible, especially for disabled people and the elderly [80].

In a dialog system of this kind, several modules cooperate to perform the interaction with the user: the Automatic Speech Recognizer (ASR), the Spoken Language Understanding Module (SLU), the Dialog Manager (DM), the Natural Language Generation module (NLG), and the Text-To-Speech Synthesizer (TTS). Each one of them has its own characteristics and the selection of the most convenient model varies depending on certain factors: the goal of each module, the possibility of manually defining the behavior of the module, or the capability of automatically obtaining models from training samples. Figure 1 shows the set of actions and main modules in the architecture of a Spoken Dialog System.

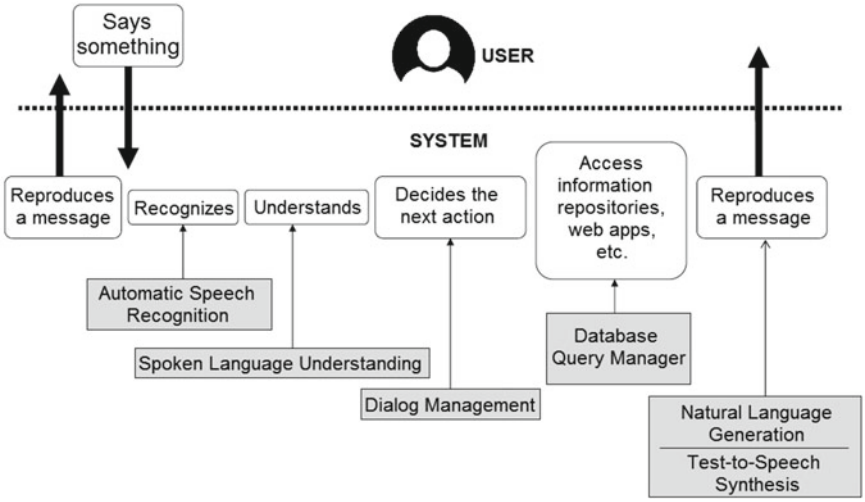

Fig. 1 Set of actions and modules in a spoken dialog system

The goal of speech recognition is to obtain the sequence of words uttered by a speaker. Once the speech recognizer has provided an output, the system must understand what the user said. The goal of spoken language understanding is to obtain the semantics from the recognized sentence. This process generally requires morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge.

The dialog manager decides the next action of the system, interpreting the incoming semantic representation of the user input in the context of the dialog. In addition, it resolves ellipsis and anaphora, evaluates the relevance and completeness of user requests, identifies and recovers from recognition and understanding errors, retrieves information from data repositories, and decides about the next system's response. Natural language generation is the process of obtaining sentences in natural language from the non-linguistic, internal representation of information handled by the dialog system. Finally, the TTS module transforms the generated sentences into synthesized speech.

In order to enable rapid deployment of these systems, markup languages such as VoiceXML[1] have been widely adopted as they reduce the time and effort required for system implementation. However, system development with this approach involves a very costly engineering cycle [62]. As an alternative, data-based models try to reduce the effort and time required to develop a new dialog system or to adapt them to deal with a new task. This kind of models are usually based on modeling the different processes probabilistically and learning the parameters of the different statistical models from a dialog corpus. This approach has been widely used for speech recognition and also for language understanding [14, 20, 37, 51, 67]. Even though in the literature there are models for dialog managers that are manually designed, over the last few years, approaches using statistical models to represent the behavior of the dialog manager have also been developed [36, 38, 74, 83].

As described by [58], there are three main categories of elements of the spoken dialog interaction where the availability of vast amounts of data (known as Big Data [2, 19, 47]) can potentially improve automation rate, and ultimately, the penetration and acceptance of speech interfaces in the wider consumer market. They are task-independent behaviors (e.g., error correction and confirmation behavior), task-specific behaviors (e.g., logic associated with certain customer-care practices), and task-interface behaviors (e.g., prompt selection). However, these three categories have in common today the lack of robust guiding principles validated by empirical evidence.

The following sections of this chapter describe the current uses of Big Data to develop conversational interfaces including speech recognition, natural language understanding, dialog management and optimization, context-awareness, emotion recognition, user adaptation and service personalization, multi-domain and multilingual services, proactiveness, and spoken language generation and synthesis.

[1] http://www.w3.org/TR/voicexml20/.

2 Spoken Language Recognition

As described in the introduction section, speech recognition is the process of obtaining the text string corresponding to an acoustic input [45, 55, 78]. It is a highly complex task, as there is a great deal of variation in input characteristics, which can differ according to the linguistics of the utterance, the speaker, the interaction context and the transmission channel. Different aspects that are usually taken into account when classifying ASR systems are the kind of users supported (user-independent or user-dependent systems), style of speech supported (recognizers isolated words, connected words or continuous speech), or vocabulary size (small, medium, or large vocabulary).

The complexity of the recognition task lies in several problems: the acoustic variability (each person pronounces sounds differently when speaking), acoustic confusion (many words sound similar), the coarticulation problem (the characteristics of spoken sounds may vary depending on neighboring sounds), out of vocabulary words and spontaneous speech (interjections, pauses, doubts, false starts, repetitions of words, self-corrections, etc.), and environmental conditions (noise, channel distortion, bandwidth limitations, etc.). For these reasons, it is very important to try to detect and correct errors generated during the ASR process, since the output of the ASR is the starting point of the other modules in a spoken dialog system.

During the last decades, the field of automatic speech recognition has progressed from the recognition of isolated words in reduced vocabularies to continuous speech recognition with increasing vocabulary sets. These advances have made the communication with dialog systems increasingly more natural. Among the variety of techniques used to develop ASR systems, the data-based approach is currently the most widely used. In this approach, the speech recognition problem can be understood as finding the word sequence W uttered by the user given a sequence of acoustic data A. This sequence can be determined by means of the following expression:

$$W = \max_{W} P(W|A) \tag{1}$$

Using the Bayes rule, the previous equation can be rewritten as follows:

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)} \tag{2}$$

where $P(A|W)$ is called the acoustic model (probability of the sequence A the word sequence W has been uttered) and $P(W)$ is provided by the language model (probabilities of sequences of words.). The probabilities of the rules in these models are learned from training data. The acoustic model is created by taking audio recordings of speech and their transcriptions and then compiling them into statistical representations of the sounds for the different words. Learning a language model requires the transcriptions of sentences related to the application domain of the system. Since

the probability of the acoustic sequence is independent of the sequence of words, the previous expression can be written as follows:

$$W = \max_{W} P(A|W)P(W) \tag{3}$$

For the practical implementation of this approach, the most widely used solution consists of modeling the acoustic units by means of Hidden Markov Models (HMM), as it is the case of speech recognizers widely used by the scientific community as HTK (Hidden Markov Model Toolkit)[2] or CMU Sphinx.[3]

The success of the HMM is mainly based on the use of machine learning algorithms to learn the parameters of the model [61], as well as in their ability to represent speech as a sequential phenomenon over time. Multiple models have been studied, such as discrete models, semicontinuous or continuous, as well as a variety of topologies models.

The language model is one of the essential components required to develop a recognizer of continuous speech. The most used language models are based on N-grams [3, 28] and regular or context-free grammars [29, 67]. Grammars are usually suitable for small tasks, providing more precision based on the type of restrictions. However, they are not able to represent the great variability of natural speech processes.

N-grams models allow to collect more easily the different concatenations among words when a sufficient number of training samples is available. In an n-gram model, the probability $P(w_1, \ldots, w_m)$ of observing the sentence $w_1, \ldots, w_m$ is approximated as

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) \tag{4}$$

This equation assumes that the probability of observing the i-th word w_i in the context history of the preceding $i-1$ words can be approximated by the probability of observing it in the shortened context history of the preceding $n-1$ words (n-th order Markov property). The conditional probability can be calculated from n-gram model frequency counts:

$$P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{\text{count}(w_{i-(n-1)}, \ldots, w_{i-1}, w_i)}{\text{count}(w_{i-(n-1)}, \ldots, w_{i-1})} \tag{5}$$

Figure 2 shows an example of the estimation of the bigram probabilities using the Maximum Likelihood Estimate.[4] Typically, the probabilities are not derived directly from the frequency counts. Instead, some form of smoothing is necessary, assigning

[2] http://htk.eng.cam.ac.uk/.

[3] http://cmusphinx.sourceforge.net/.

[4] https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf.

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$
$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

Fig. 2 Estimating bigram probabilities by means of the maximum likelihood estimate

some of the total probability mass to unseen words or n-grams. Various methods are then used, from simple "add-one" smoothing (assign a count of 1 to unseen n-grams) to more sophisticated models, such as Good-Turing discounting or back-off models.

From around 2010, Deep Neural Networks (DNNs) have replaced HMM models. DNNs are now used extensively in industrial and academic research as well as in most commercially deployed ASR systems. Various studies have shown that DNNs outperform HMM models in terms of increased recognition accuracy [24, 68]. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. As described in [53], a key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabeled and uncategorized.

3 Spoken Language Understanding

Once the spoken dialog system has recognized what the user uttered, it is necessary to understand what he said [46, 50, 85]. Natural language processing is a method of obtaining the semantics of a text string and generally involves morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge. In the first stage, lexical and morphological knowledge divide the words into their constituents by distinguishing between lexemes and morphemes: lexemes are parts of words that indicate their semantics and morphemes are the different infixes and suffixes that provide different word classes.

Syntactic analysis yields the hierarchical structure of the sentences. However, in spoken language, phrases are frequently affected by difficulties associated with the so-called disfluency phenomena: filled pauses, repetitions, syntactic incompleteness and repairs [18]. Semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituent parts. In the pragmatic and discourse-processing stage, the sentences are interpreted in the context of the whole dialog, the main complexity of this stage is the resolution of anaphora, and ambiguities derived from phenomena such as irony, sarcasm or double entendre.

The process of understanding can be understood as a change in language representation, from natural language to a semantic language, so that the meaning of the message is not changed. As in the speech recognizer, the spoken language understanding module can work with several hypotheses (both for recognition and understanding) and confidence measures. There are currently two major approaches to tackling the problem of understanding: rule-based approaches and statistical models learned from data corpus.

Rule-based approaches extract semantic information based on a syntactic-semantic analysis of the sentences, using grammars defined for the task, or by means of the detection of keywords with semantic meanings. Some analyzers, in order to improve the robustness of the analysis, combine syntactic and semantic aspects of the specific task. Other techniques are based on an analysis at two levels, in which grammars are used to carry out a detailed analysis of the sentence and extract relevant semantic information. In addition, there are systems that use rule-based analyzers automatically learned from a training corpus using natural language processing techniques.

In the case of statistical methods, the process is based on the definition of linguistic units with semantic content and obtaining models from labeled samples. This type of analysis [50, 67] uses a probabilistic model to identify concepts, markers and values of cases, to represent the relationship between markers of cases and their values and to decode semantically pronunciations of the user. The model is generated during a training phase (learning), where its parameters capture the correspondences between text entries and semantic representation. Once the training model has been learned, it is used as a decoder to generate the best representation.

The semantic definition is usually based on the concept of frame in most of the current dialog systems. In this approach, the representation generated by the spoken language understanding module contains concepts (different types of queries that users can require to the system) and attributes (information to be provided by the user to complete or modify the queries). Thus, every message sent by the spoken language understanding module to the dialog manager after each user utterance consists of a frame structure.

4 Dialog Management

Although dialog management is only a part of the development cycle of spoken dialog systems, it can be considered one of the most demanding tasks given that this module encapsulates the logic of the speech application [81]. [77] state that dialog management involves four main tasks: (i) updating the dialog context, (ii) providing a context for sentence interpretation, (iii) coordinating other modules and (iv) deciding the information to convey to the user and when to do it. Thus, the selection of a specific system action depends on multiple factors, such as the output of the speech recognizer (e.g., measures that define the reliability of the recognized information), the dialog interaction and previous dialog history (e.g., the number of

repairs carried out so far), the application domain (e.g., guidelines for customer service), knowledge about the users, and the responses and status of external back-ends, devices, and data repositories. Given that the actions of the system directly impact users, the dialog manager is largely responsible for user satisfaction. This way, the design of an appropriate dialog management strategy is at the core of dialog system engineering.

Statistical approaches for dialog management present several important advantages with regard traditional rule-based methodologies. Rather than maintaining a single hypothesis for the dialog state, they maintain a distribution over many hypotheses for the correct dialog state. In addition, statistical methodologies choose actions using an optimization process, in which a developer specifies high-level goals and the optimization works out the detailed dialog plan. For instance, Hoxha and Weng have very recently proposed a mixed-initiative dialog-based approach to support autonomous clinical data access and recommend needed technology development and communication study for accelerating clinical research [26].

Automating dialog management is useful for developing, deploying and re-deploying applications and also reducing the time-consuming process of handcrafted design. In fact, the application of machine learning approaches to dialog management strategy design is a rapidly growing research area. Machine-learning approaches to dialog management attempt to learn optimal strategies from corpora of real human-computer dialog data using automated "trial-and-error" methods instead of relying on empirical design principles [86]. The main trend in this area is an increased use of data for automatically improving the performance of the system.

Statistical models can be trained with corpora of human-computer dialogs with the goal of explicitly modeling the variance in user behavior that can be difficult to address by means of hand-written rules [66]. Additionally, if it is necessary to satisfy certain deterministic behaviors, it is possible to extend the strategy learned from the training corpus with handcrafted rules that include expert knowledge or specifications about the task [32, 72, 75, 87].

The goal is to build systems that exhibit more robust performance, improved portability, better scalability and easier adaptation to other tasks. However, model construction and parameterization is dependent on expert knowledge, and the success of statistical approaches is dependent on the quality and coverage of the models and data used for training [66]. Moreover, the training data must be correctly labeled for the learning process. The size of currently available annotated dialog corpora is usually too small to sufficiently explore the vast space of possible dialog states and strategies. Collecting a corpus with real users and annotating it requires considerable time and effort.

To address these problems, researchers have proposed alternative techniques that facilitate the acquisition and labeling of corpora, such as Wizard of Oz [16, 31], bootstrapping [1, 15], active learning [9, 41], automatic dialog act classification and labeling [56, 79], and user simulation [43, 66].

Another relevant problem is how to deal with unseen situations, that is, situations that may occur during the dialog and that were not considered during training. To address this point it is necessary to employ generalizable models in order to obtain appropriate system responses that enable to continue with the dialog in a satisfactory way.

Another difficulty is in the design of a good dialog strategy, which in many cases is far from being trivial. In fact, there is no clear definition of what constitutes a good dialog strategy [35, 66]. Users are diverse, which makes it difficult to foresee which form of system behavior will lead to quick and successful dialog completion, and speech recognition errors may introduce uncertainty about their intention.

The most widespread methodology for machine-learning of dialog strategies consists of modeling human-computer interaction as an optimization problem using Markov Decision Processes (MDP) and reinforcement methods [38, 70]. The main drawback of this approach is that the large state space of practical spoken dialog systems makes its direct representation intractable [89]. Partially Observable MDPs (POMDPs) outperform MDP-based dialog strategies since they provide an explicit representation of uncertainty [63]. This enables the dialog manager to avoid and recover from recognition errors by sharing and shifting probability mass between multiple hypotheses of the current dialog state.

An approach that scales the POMDP framework for implementing practical spoken dialog systems by the definition of two state spaces is presented in [88]. Approximate algorithms have also been developed to overcome the intractability of exact algorithms but even the most efficient of these techniques such as Point-Based Value Iteration (PBVI) cannot scale to the many thousand states required by a statistical dialog manager [82]. Composite Summary Point-Based Value Iteration (CSPBVI) has suggested the use of a small summary space for each slot where PBVI policy optimization can be applied. However, policy learning in this technique can only be performed offline, i.e. at design time, because policy training requires an existing accurate model of user behavior. An alternative technique for online training based on Q-learning is presented in [73], which allows the system to adapt to real users as new dialogs are recorded. This technique does not require any model of user behavior, so user simulation techniques are proposed to iteratively learn the dialog model.

Other authors have combined conventional dialog managers with a fully-observable Markov decision process [21, 71], or proposed using multiple POMDPs and selecting actions using hand-crafted rules [82]. In [84], the authors combine the robustness of the POMDP with the developer control afforded in conventional approaches: the (conventional) dialog manager and POMDP run in parallel, but the dialog manager is augmented so that it outputs one or more allowed actions at each time-step. The POMDP then chooses the best action from this limited set. Results from a real voice dialer application show that adding the POMDP machinery to a standard dialog system can yield a significant improvement [84].

Other interesting approaches for statistical dialog management are based on modeling the system by means of Hidden Markov Models [10], stochastic Finite-State Transducers [25, 27, 60], or using Bayesian Networks [49, 57]. Also [33] proposed

a different hybrid approach to dialog modeling in which n-best recognition hypotheses are weighted using a mixture of expert knowledge and data-driven measures by using an agenda and an example-based machine translation approach respectively.

5 Natural Language Generation

Natural language generation is the process of obtaining texts in natural language from a non-linguistic representation [34, 42]. It is usually carried out in 5 steps: content organization, content distribution in sentences, lexicalization, generation of referential expressions and linguistic realization. It is important to obtain legible messages, optimizing the text using referring expressions and linking words and adapting the vocabulary and the complexity of the syntactic structures to the users linguistic expertise.

The simplest approach consists of using predefined text messages (e.g. error messages and warnings). Although intuitive, this approach completely lacks from flexibility. The next level of sophistication is template-based generation, in which the same message structure is produced with slight alterations. The template approach is used mainly for multi-sentence generation, particularly in applications whose texts are fairly regular in structure, such as business reports.

Phrase-based systems employ what can be considered as generalized templates at the sentence level (in which case the phrases resemble phrase structure grammar rules), or at the discourse level (in which case they are often called text plans). In such systems, a pattern is first selected to match the top level of the input, and then each part of the pattern is expanded into a more specific one that matches some portion of the input. The cascading process stops when every pattern has been replaced by one or more words.

Finally, feature-based systems represent the maximum level of generalization and flexibility. In feature-based systems, each possible minimal alternative of expression is represented by a single feature; for example, whether the sentence is either positive or negative, if it is a question or an imperative or a statement, or its tense. To arrange the features it is necessary to employ linguistic knowledge. Another alternative is to use corpus-based natural language generation [54], which stochastically generates system utterances.

6 Text-To-Speech Synthesis

Text-to-speech synthesizers transform a text into an acoustic signal [11]. A text-to-speech system is composed of two parts: a front-end and a back-end. The front-end carries out two major tasks. Firstly, it converts raw text containing symbols such as numbers and abbreviations into their equivalent words. This process is often called text normalization, pre-processing, or tokenization. Secondly, it assigns a phonetic

transcriptions to each word, and divides and marks the text into prosodic units, i.e. phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. The output of the front-end is the symbolic representation constituted by the phonetic transcriptions and prosody information.

The back-end (often referred to as the synthesizer) converts the symbolic linguistic representation into sound. On the one hand, speech synthesis can be based on human speech production. This is the case of parametric synthesis which simulates the physiological parameters of the vocal tract, and formant-based synthesis, which models the vibration of vocal chords. In this technique, parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. Another approach based on physiological models is articulatory synthesis, which refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes.

On the other hand, concatenative synthesis employs pre-recorded units of human voice. Concatenative synthesis is based on stringing together segments of recorded speech. It generally produces the most natural-sounding synthesized speech; however, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. The quality of the synthesized speech depends on the size of the synthesis unit employed.

Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Unit selection provides the greatest naturalness, because it applies only a small amount of digital signal processing to the recorded speech. There is a balance between intelligibility and naturalness of the voice output or the automatization of the synthesis procedure. For example, synthesis based on whole words is more intelligible than the phone-based but for each new word it is necessary to obtain a new recording, whereas the phones allow building any new word. In one extreme, domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. It is used in applications in which the variety of texts the system will produce is limited to a particular domain, like transit schedule announcements or weather reports.

At the other extreme, diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones and German about 2,500. In diphone synthesis, only one example of each diphone is contained in the speech database. Finally, HMM-based synthesis is a method in which the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves, based on the maximum likelihood criterion.

7 User Modeling and Evaluation of the System

Research in techniques for user modeling has a long history within the fields of language processing and conversational agents. The main purpose of a simulated user in this field is to improve the usability of a conversational agent through the generation of corpora of interactions between the system and simulated users [52], reducing time and effort required for collecting large samples of interactions with real users. Moreover, each time changes are made to the system it is necessary to collect more data in order to evaluate the changes. Thus, the availability of large corpora acquired with a user simulator should contribute positively to the development of the system.

User simulators can be used to evaluate different aspects of a conversational agent, particularly at the earlier stages of development, or to determine the effects of changes to the system's functionalities (e.g., evaluate confirmation strategies or introduce of errors or unpredicted answers in order to evaluate the capacity of the dialog manager to react to unexpected situations). A second usage, in which we are mainly interested in this contribution, is to support the automatic learning of optimal dialog strategies using statistical methodologies. Large amounts of data are required for a systematic exploration of the dialog state space and corpora acquired with simulated users are extremely valuable for this purpose.

Two main approaches can be distinguished to the creation of simulated users: rule based and data or corpus based. In a rule-based simulated user the researcher can create different rules that determine the behavior of the system [8, 39, 44]. This approach is particularly useful when the purpose of the research is to evaluate the effects of different dialog management strategies. In this way the researcher has complete control over the design of the evaluation study.

Data-based user models are based on probabilistic methods to generate the user input, with the advantage that this uncertainty can better reflect the unexpected behaviors of users interacting with the system. Statistical models for modeling user behavior have been suggested as the solution to the lack of the data that is required for training and evaluating dialog strategies. Using this approach, the dialog manager can explore the space of possible dialog situations and learn new potentially better strategies. Methodologies based on learning user intentions have the purpose of optimizing dialog strategies. A summary of user simulation techniques for reinforcement learning of the dialog strategy can be found in [66].

The most extended methodology for machine-learning of dialog strategies consists of modeling human-computer interaction as an optimization problem using Markov Decision Process (MDP) and reinforcement methods [38]. The main drawback of this approach is the large state space of practical spoken dialog systems, whose representation is intractable if represented directly. Although Partially Observable MDPs (POMDPs) outperform MDP-based dialog strategies, they are limited to small-scale problems, since the state space would be huge and exact POMDP optimization is again intractable [83].

In [12, 13], Eckert, Levin and Pieraccini introduced the use of statistical models to predict the next user action by means of a n-gram model. The proposed model has the advantage of being both statistical and task-independent. Its weak point consists of approximating the complete history of the dialog by a bigram model. In [38], the bigram model is modified by considering only a set of possible user answers following a given system action (the Levin model). Both models have the drawback of considering that every user response depends only on the previous system turn. Therefore, the simulated user can change objectives continuously or repeat information previously provided.

Georgila, Henderson and Lemon propose the use of HMMs, defining a more detailed description of the states and considering an extended representation of the history of the dialog [17]. Dialog is described as a sequence of *Information States* [7]. Two different methodologies are described to select the next user action given a history of information states. The first method uses n-grams [12], but with values of n from 2 to 5 to consider a longer history of the dialog. The best results are obtained with 4-grams. The second methodology is based on the use of a linear combination of 290 characteristics to calculate the probability of every action for a specific state.

Cuayáhuitl et al. present a method for dialog simulation based on HMMs in which both user and system behaviors are simulated [10]. Instead of training only a generic HMM model to simulate any type of dialog, the dialogs of an initial corpus are grouped according to the different objectives. A submodel is trained for each one of the objectives, and a bigram model is used to predict the sequence of objectives.

In [64], a new technique for user simulation based on explicit representations of the user goal and the user agenda is presented. The user agenda is a structure that contains the pending user dialog acts that are needed to elicit the information specified in the goal. This model formalizes human-machine dialogs at a semantic level as a sequence of states and dialog acts. An EM-based algorithm is used to estimate optimal parameter values iteratively. In [65], the agenda-based simulator is used to train a statistical POMDP-based dialog manager.

A data-driven user intention simulation method that integrates diverse user discourse knowledge (cooperative, corrective, and self-directing) is presented in [30]. User intention is modeled based on logistic regression and Markov logic framework. Human dialog knowledge is designed into two layers, domain and discourse knowledge, and integrated with the data-driven model in generation time. A methodology of user simulation applied to the evaluation and refinement of stochastic dialog systems is presented in [76]. The proposed user simulator incorporates several knowledge sources, combining statistical and heuristic information to enhance the dialog models by an automatic strategy learning. As it is described in the following section, our proposed user simulation technique is based on a classification process that considers the complete dialog history by incorporating several knowledge sources, combining statistical and heuristic information to enhance dialog models by an automatic strategy learning.

In the area of user modeling and dialog systems, emotion has been used for several purposes, as summarized in the taxonomy of applications proposed in [5]. In some application domains, it is fundamental to recognize the affective state of the user to

adapt the systems behavior. For example, in emergency services [6] or intelligent tutors [40], it is necessary to know the user emotional state to calm them down, or to encourage them in learning activities. For other applications domains, it can also play an important role in order to solve stages of the dialog that cause negative emotional states, avoid them and foster positive ones in future interactions. Bain have recently presented a proposal to extract emotional information using cloud-based Big Data infrastructure and mobile devices [4]. Hosain et al. has also very recently proposed an infrastructure that combines the potential of emotion-aware Big Data and cloud technology towards the future generation mobile communication technologies (5G) [69].

8 Future Research and Challenges

Throughout the last years, some experts have dared to envision what the future research guidelines in the application of multimodal dialog systems for educative purposes would be based on the advances in Big Data research. These objectives have gradually changed towards ever more complex goals, such as providing the system with advanced reasoning, problem solving capabilities, adaptiveness, proactiveness, affective intelligence, and multilinguality. All these concepts are not mutually exclusive, as for example the system's intelligence can also be involved in the degree to which it can adapt to new situations, and this adaptiveness can result in better portability for use in different environments.

As can be observed, these new objectives refer to the system as a whole, and represent major trends that in practice are achieved through joint work in different areas and components of the dialog system. Thus, current research trends are characterized by large-scale objectives which are shared out between the different researchers in different areas.

Proactiveness is necessary for computers to stop being considered a tool and becoming real conversational partners. Proactive systems have the capability of engaging in a conversation with the user even when he has not explicitly requested the system's intervention. This is a key aspect in the development of ubiquitous computing architectures in which the system is embedded in the user's environment, and thus the user is not aware that he is interacting with a computer, but rather he perceives he is interacting with the environment. To achieve this goal, it is necessary to provide the systems with problem-solving capabilities and context-awareness.

Adaptivity may also refer to other aspects in speech applications. There are different levels in which the system can adapt to the user. The simplest one is through personal profiles in which the users have static choices to customize the interaction. Systems can also adapt to the users' environment, for example ambient intelligence applications such as the ubiquitous proactive systems described. A more sophisticated approach is to adapt to the user's knowledge and expertise. This is especially important in educative systems to adapt the system taking into account the specific

evolution of each of the students, the previous uses of the system, and the errors that they have made during the previous interactions.

There is also an increasing interest in the development of multimodal conversational systems that dynamically adapt their conversational behaviors to the users' affective state. The empathetic educative agent can thus indeed contribute to a more positive perception of the interaction.

Portability is currently addressed from very different perspectives, the three main ones being domain, language and technological independence. Ideally, systems should be able to work over different educative application domains, or at least be easily adaptable between them. Current studies on domain independence center on how to merge lexical, syntactic and semantic structures from different contexts and how to develop dialog managers that deal with different domains.

Finally, technological independence deals with the possibility of using multimodal systems with different hardware configurations. Computer processing power will continue to increase, with lower costs for both processor and memory components. The systems that support even the most sophisticated multimodal applications will move from centralized architectures to distributed configurations and thus must be able to work with different underlying technologies.

9 Conclusions

Dialog systems appeared as a technology aimed at sustaining conversations with their users that could be considered natural and human-like. However, to achieve this long pursued objective, these systems must be able to operate in a wide range of domains and tasks, some of them difficult to process and complex [19], for which being able to learn from massive amounts of data becomes crucial to show appropriate behaviors.

We have addressed Big Data as (1) new sources of huge amounts of data, and (2) as the novel machine learning and information extraction algorithms that have appeared to process them. On the one hand, web pages and searches, social network, blog posts, and emails, they all provide an invaluable source for natural language resources. Similarly, voice calls, recorded dialogs and conversations have a huge potential to provide insights into human conversational behavior. However, manually examining such Big Data is laborious and error-prone. On the other hand, the emergence of different statistical approaches has enabled to accurately analyze unstructured data with a double benefit: to be less dependent on intensive manual annotation, and to gain a better understanding of human conversation by learning more accurate models from a more representative amount of data.

In this chapter we have discussed the tremendous potential of Big Data to improve several aspects of dialog system research and development, including speech processing, natural language understanding and dialog management.

References

1. Abdennadher S, Aly M, Bhler D, Minker W, Pittermann J (2007) Becam tool - a semi-automatic tool for bootstrapping emotion corpus annotation and management. In: Proceedings of the international conference on spoken language processing (Interspeech'2007), pp 946–949
2. Agerri R, Artola X, Beloki Z, Rigau G, Soroa A (2015) Big data for natural language processing: a streaming approach. Knowl-Based Syst 79:36–42
3. Bahl L, Jelinek F, Mercer R (1990) A maximum likelihood approach to continuous speech recognition. Readings in Speech recognition, pp 308–319
4. Baimbetov Y, Khalil I, Steinbauer M, Anderst-Kotsis G (2015) Using Big Data for emotionally intelligent mobile services through multi-modal emotion recognition. Springer, pp 127–138
5. Batliner A, Burkhardt F, van Ballegooy M, Noth E (2006) A taxonomy of applications that utilize emotional awareness. In: Proceedings of 1st international language technologies conference (IS-LTC 06), pp 246–250
6. Bickmore T, Giorgino T (2004) Some novel aspects of health communication from a dialogue systems perspective. In: Proceedings of AAAI fall symposium on dialogue systems for health communication, pp 275–291
7. Bos J, Klein E, Lemon O, Oka T (2003) DIPPER: description and formalisation of an information-state update dialogue system architecture. In: Proceedings of the SIGdial, pp 115–124
8. Chung G (2004) Developing a flexible spoken dialog system using simulation. In: Proceedings of ACL, pp 63–70
9. Cohn DA, Atlas L, Ladner R (1994) Improving generalization with active learning. Mach Learn 15(2):201–221
10. Cuayhuitl H, Renals S, Lemon O, Shimodaira H (2005) Human-computer dialogue simulation using hidden Markov models. In: Proceedings of ASRU, pp 290–295
11. Dutoit T (1996) An introduction to text-to-speech synthesis. Kluwer Academic Publishers
12. Eckert W, Levin E, Pieraccini R (1997) User modeling for spoken dialogue system evaluation. In: Proceedings of ASRU, pp 80–87
13. Eckert W, Levin E, Pieraccini R (1998) Automatic evaluation of spoken dialogue systems. Technical report, TR98.9.1, ATT Labs Research
14. Esteve Y, Raymond C, Bechet F, Mori RD (2003) Conceptual decoding for spoken dialog systems. In: Proceedings of European conference on speech communications and technology (Eurospeech'03). vol 1, pp 617–620
15. Fabbrizio GD, Tur G, Hakkani-Tr D, Gilbert M, Renger B, Gibbon D, Liu Z, Shahraray B (2008) Bootstrapping spoken dialogue systems by exploiting reusable libraries. Nat Lang Eng 14(3):313–335
16. Fraser M, Gilbert G (1991) Simulating speech systems. Comput Speech Lang 5:81–99
17. Georgila K, Henderson J, Lemon O (2005) Learning user simulations for information state update dialogue systems. In: Proceedings of Eurospeech'05, pp. 893–896
18. Gibbon D, Mertins I (Eds.), R.M.: Handbook of multimodal and spoken dialogue systems: resources, terminology and product evaluation. Kluwer Academic Publishers (2000)
19. Gudivada VN, Rao D, Raghavan VV (2015) Big data driven natural language processing research and applications, vol 33. Elsevier, pp 203–238
20. He Y, Young S (2003) A data-driven spoken language understanding system. In: Proceedings of IEEE Automatic speech recognition and understanding workshop (ASRU'03), pp 583–588
21. Heeman P (2007) Combining reinforcement learning with information-state update rules. In: Proceedings of the 8th Annual conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07), pp 268–275
22. Heinroth T, Minker W (2012) Introducing spoken dialogue systems into intelligent environments. Kluwer Academic Publishers, Springer
23. Hempel T (2008) Usability of speech dialog systems: listening to the target audience. Springer

24. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
25. Hori C, Ohtake K, Misu T, Kashioka H, Nakamura S (2009) Recent advances in WFST-based dialog system. In: Proceedings of the international conference on spoken language processing (Interspeech'2009), pp 268–271
26. Hoxha J, Weng C (2016) Leveraging dialog systems research to assist biomedical researchers interrogation of big clinical data. J Biomed Inf 61:176–184
27. Hurtado L, Planells J, Segarra E, Sanchis E, Griol D (2010) A stochastic finite-state transducer approach to spoken dialog management. In: Proceedings of the international conference on spoken language processing (Interspeech'2010), pp 3002–3005
28. Jelinek F (1990) Self-organized language modeling for speech recognition. Readings in Speech recognition, pp 450–506
29. Jelinek F, Lafferty MR (1992) Basic methods of probabilistic context free grammars. Springer, pp 345–360
30. Jung S, Lee C, Kim K, Lee D, Lee G (2011) Hybrid user intention modeling to diversify dialog simulations. Comput Speech Lang 25(2):307–326
31. Lane I, Ueno S, Kawahara T (2004) Cooperative dialogue planning with user and situation models via example-based training. In: Proceedings of workshop on man-machine symbiotic systems, pp 2837–2840, Kyoto, Japan
32. Laroche R, Putois G, Bretier P, Young S, Lemon O (2008) Requirements analysis and theory for statistical learning approaches in automaton-based dialogue management. Technical report, School of Informatics, Edinburgh University, Edinburgh, UK
33. Lee C, Jung S, Kim K, Lee GG (2010) Hybrid approach to robust dialog management using agenda and dialog examples. Comput Speech Lang 24(4):609–631
34. Lemon O (2011) Learning what to say and how to say it: joint optimisation of spoken dialogue management and natural language generation. Comput Speech Lang 25(2):210–221
35. Lemon O, Pietquin O (2012) Data-Driven methods for adaptive spoken dialogue systems. Computational learning for conversational interfaces. Springer, Berlin
36. Lemon O, Georgila K, Henderson J (2006) Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users: the TALK TownInfo evaluation. In: Proceedings of IEEE-ACL workshop on spoken language technology (SLT'06), pp 178–181
37. Levin E, Pieraccini R (1995) Concept-based spontaneous speech understanding system. In: Proceedings of European conference on speech communications and technology (Eurospeech'95). pp. 555–558 (1995)
38. Levin E, Pieraccini R, Eckert W (2000) A stochastic model of human-machine interaction for learning dialog strategies. IEEE Trans Speech Audio Process 8(1):11–23
39. Lin B, Lee L (2001) Computer aided analysis and design for spoken dialogue systems based on quantitative simulations. IEEE Trans Speech Audio Process 9(5):534–548
40. Litman D, Forbes-Riley K (2006) Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun 48(5):559–590
41. Liu Y, Shriberg E (2005) Does active learning help automatic dialog act tagging in meeting data. In: Proceedings of the international conference on spoken language processing (Interspeech'2005), pp 2777–2780, Lisbon, Portugal
42. López V, Eisman E, Castro J, Zurita J (2011) A case based reasoning model for multilingual language generation in dialogues. Expert Syst Appl 39(8):7330–7337
43. López-Cózar R, Callejas Z, McTear M (2006) Testing the performance of spoken dialogue systems by means of an artificially simulated user. Artif Intell Rev 26:291–323
44. López-Cózar R, la Torre AD, Segura J, Rubio A, Sánchez V (2003) Assessment of dialogue systems by means of a new simulation technique. Speech Commun 40(3):387–407
45. López-Cózar R, Callejas Z (2008) ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information. Comput Speech Lang 50(8–9):745–766

46. López-Cózar R, Callejas Z, Griol D (2010) ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information. Knowl-Based Syst 23(5):471–485
47. Mayer-Schonberger V (2003) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan-Houghton Mifflin Harcourt
48. McTear MF, Callejas Z, Griol D (2016) The conversational interface. Springer
49. Meng HH, Wai C, Pieraccini R (2003) The use of belief networks for mixed-initiative dialog modeling. IEEE Trans Speech Audio Process 11(6):757–773
50. Minker W (1999) Design considerations for knowledge source representations of a stochastically-based natural language understanding component. Speech Commun 28(2):141–154
51. Minker W, Waibel A, Mariani J (1999) Stochastically-based semantic analysis. Kluwer Academic Publishers, Dordrecht (Holland)
52. Mller S, Englert R, Engelbrecht K, Hafner V, Jameson A, Oulasvirta A, Raake A, Reithinger N (2006) MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In: Proceedings of the Interspeech, pp 1786–1789
53. Najafabadi M, Villanustre F, Khoshgoftaar T, Seliya N, WaldEmail R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1)
54. Oh AH, Rudnicky AI (2000) Stochastic language generation for spoken dialogue systems. In: Proceedings of ANLP/NAACL workshop on conversational systems, pp 27–32
55. O'Shaughnessy D (2008) Automatic speech recognition: history, methods and challenges. Pattern Recogn 41(10):2965–2979
56. O'Shea J, Bandar Z, Crockett K (2012) A multi-classifier approach to dialogue act classification using function words. Lecture notes in computer science, vol 7270, pp 119–143
57. Paek T, Horvitz E (2000) Conversation as action under uncertainty. In: Proceedings of the 16th conference on uncertainty in artificial intelligence, pp 455–464
58. Paek T, Pieraccini R (2008) Automating spoken dialogue management design using machine learning: an industry perspective. Speech Commun 50(8–9):716–729
59. Pieraccini R (2012) The voice in the machine: building computers that understand speech. MIT Press
60. Planells J, Hurtado L, Sanchis E, Segarra E (2012) An online generated transducer to increase dialog manager coverage. In: Proceedings of the international conference on spoken language processing (Interspeech'2012)
61. Rabiner L, Juang B, Lee C (1996) An overview of automatic speech recognition. Kluwer Academic Publishers, pp 1–30
62. Rojas-Barahona L, Giorgino T (2009) Adaptable dialog architecture and runtime engine (adarte): a framework for rapid prototyping of health dialog systems. Int J Med Inf 78:56–68
63. Roy N, Pineau J, Thrun S (2000) Spoken dialogue management using probabilistic reasoning. In: Proceedings of the 38th Annual meeting of the association for computational linguistics (ACL'00), pp 93–100
64. Schatzmann J, Thomson B, Weilhammer K, Ye H, Young S (2007) Agenda-based user simulation for bootstrapping a POMDP dialogue system. In: Proceedings of HLT/NAACL, pp 149–152
65. Schatzmann J, Thomson B, Young S (2007) Statistical user simulation with a hidden agenda. In: Proceedings of SIGdial, pp 273–282
66. Schatzmann J, Weilhammer K, Stuttle M, Young S (2006) A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowl Eng Rev 21(2):97–126
67. Segarra E et al (2002) Extracting semantic information through automatic learning techniques. Int J Pattern Recogn Artif Intell 16(3):301–307
68. Seide F, Li G, Yu D (2011) Conversational speech transcription using context-dependent deep neural networks. In: Proceedings of the 12th annual conference of the international speech communication association (InterSpeech 2011), pp 437–440. Florence, Italy

69. Shamim-Hossain M, Muhammad G, Alhamid MF, Song B, Al-Mutib K (2016) Audio-visual emotion recognition using big data towards 5G. Mobile Netw Appl 1:1–11
70. Singh S, Kearns M, Litman D, Walker M (1999) Reinforcement learning for spoken dialogue systems. In: Proceedings of neural information processing systems (NIPS'99), pp 956–962
71. Singh S, Litman D, Kearns M, Walker M (2002) Optimizing dialogue management with reinforcement leaning: experiments with the NJFun system. J Artif Intell 16:105–133
72. Suendermann D, Pieraccini R (2012) One year of contender: what have we learned about assessing and tuning industrial spoken dialog systems? In: Proceedings of NAACL-HLT workshop on future directions and needs in the spoken dialog community: tools and data (SDCTD'12), pp 45–48
73. Thomson B, Schatzmann J, Weilhammer K, Ye H, Young S (2007) Training a real-world POMDP-based Dialog System. In: Proceedings of NAACL-HLT-Dialog'07 workshop on bridging the gap: academic and industrial research in dialog technologies, pp 9–16
74. Torres F, Sanchis E, Segarra E (2003) Development of a stochastic dialog manager driven by semantics. In: Proceedings of European conference on speech communications and technology (Eurospeech'03), pp 605–608
75. Torres F, Sanchis E, Segarra E (2008) User simulation in a stochastic dialog system. Comput Speech Lang 22:230–255
76. Torres F, Sanchis E, Segarra E (2008) User simulation in a stochastic dialog system. Comput Speech Lang 22(3):230–255
77. Traum D, Larsson S (2003) The information state approach to dialogue management. Kluwer, pp 325–353
78. Tsilfidis A, Mporas I, Mourjopoulos J, Fakotakis N (2013) Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing. Comput Speech Lang 27(1):380–395
79. Venkataraman A, Stolcke A, Shriberg E (2002) Automatic dialog act labeling with minimal supervision. In: Proceedings of the 9th Australian international conference on speech science & technology
80. Vipperla R, Wolters M, Renals S (2012) Spoken dialogue interfaces for older people. IOS Press, pp 118–137
81. Wilks Y, Catizone R, Worgan S, Turunen M (2011) Some background on dialogue management and conversational speech for dialogue systems. Comput Speech Lang 25:128–139
82. Williams J, Poupart P, Young S (2006) Partially Observable Markov decision processes with continuous observations for dialogue management. Springer, pp 191–217
83. Williams J, Young S (2007) Partially observable Markov decision processes for spoken dialog systems. Comput Speech Lang 21(2):393–422
84. Williams J (2009) The best of both worlds: unifying conventional dialog systems and pomdps. In: Proceedings of Interspeech, pp 1173–1176
85. Wu WL, Lu RZ, Duan JY, Liu H, Gao F, Chen YQ (2010) Spoken language understanding using weakly supervised learning. Comput Speech Lang 24(2):358–382
86. Young S (2002) The statistical approach to the design of spoken dialogue systems. Technical report, CUED/F-INFENG/TR.433, Cambridge University Engineering Department, Cambridge, UK
87. Young S, Gasic M, Thomson B, Williams J (2013) Pomdp-based statistical spoken dialogue systems: a review. In: Proceedings of the IEEE, pp 1–18, Montreal, Canada
88. Young S, Williams J, Schatzmann J, Stuttle M, Weilhammer K (2005) The hidden information state approach to dialogue management. Technical report, Department of Engineering, University of Cambridge, Cambridge, UK
89. Young S, Schatzmann J, Weilhammer K, Ye H (2007) The hidden information state approach to dialogue management. In: Proceedings of the 32nd IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 149–152