Measuring the differences between human-human and human-machine dialogs

David Griol and José Manuel Molina
Applied Artificial Intelligence Group (GIAA), Computer Science Department, Carlos III University of Madrid, Avda. de la Universidad - 30, 28911 - Leganés (Spain).

KEYWORD

Conversational Agents; Spoken Interaction; Spoken Dialog Systems; User Modeling; Dialog simulation; Dialog Structure Annotation.

ABSTRACT

In this paper, we assess the applicability of user simulation techniques to generate dialogs which are similar to real human-machine spoken interactions. To do so, we present the results of the comparison between three corpora acquired by means of different techniques. The first corpus was acquired with real users. A statistical user simulation technique has been applied to the same task to acquire the second corpus. In this technique, the next user answer is selected by means of a classification process that takes into account the previous dialog history, the lexical information in the clause, and the subtask of the dialog to which it contributes. Finally, a dialog simulation technique has been developed for the acquisition of the third corpus. This technique uses a random selection of the user and system turns, defining stop conditions for automatically deciding if the simulated dialog is successful or not. We use several evaluation measures proposed in previous research to compare between our three acquired corpora, and then discuss the similarities and differences with regard to these measures.

1. Introduction

Conversational agents have currently became a strong alternative to provide computers with intelligent and natural communicative capabilities. A conversational agent is a software that accepts natural language as input and generates natural language as output, engaging in a conversation with the user (Pieraccini, 2012; Heinroth and Minker, 2012; Pérez-Marín and Pascual-Nieto, 2011). To successfully interact with the users, spoken conversational agents usually carry out five main tasks: automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS).

Spoken interaction can be the only way to access information in some cases, like for example when the screen is too small to display information (e.g. hand-held devices) or when the eyes of the user are busy in other tasks (e.g. driving) (Weng et al., 2006). It is also useful for remote control of devices and robots, specially in smart environments (Menezes et al., 2007). One of the most wide-spread applications is information retrieval. Some sample applications are tourist and travel information (Glass et al., 1995), weather forecast over the phone (Zue et al., 2000), speech controlled banking systems (Melin et al., 2001), conference help (Bohus et al., 2007), etc. They have also been used for education and training, particularly in improving phonetic and linguistic skills: assistance and guidance to F18 aircraft personnel during maintenance tasks (Bohus and Rudnicky, 2002), dialog applications for computer-aided speech therapy with different language pathologies (Vaquero et al., 2006). Finally, one of the most demanding applications for fully natural and understandable dialogs, are embodied conversational agents and companions (Brahnam, 2009; Bailly et al., 2010).

The application of statistical approaches to dialog management in dialog systems has reached a growing

interest during the last decade (Schatzmann et al., 2006; Griol et al., 2014). Statistical models can be trained from real dialogs, modeling the variability in user behaviors. Although the construction and parameterization of the model depend on the expert knowledge of the task, the final objective is to develop dialog systems that have a more robust behavior, better portability, and are easier to adapt to different user profiles or tasks in Ambient Intelligence applications.

The success of statistical approaches depends on the quality of the data used to develop the dialog model. Considerable effort is necessary to acquire and label a corpus with the data necessary to train a good model. A technique that has attracted increasing interest in the last decade is based on the automatic generation of dialogs between the dialog manager and an additional module, called *user simulator*, which represents user interactions with the dialog system. The user simulator makes it possible to generate a large number of dialogs in a very simple way. Therefore, this technique reduces the time and effort that would be needed for the evaluation of a dialog system each time the system is modified.

In the literature, there are several corpus-based approaches for developing user simulators, learning optimal management strategies, and evaluating the dialog system (Schatzmann et al., 2006). Different studies have been carried out to compare the corpora acquired by means of different techniques and to define the most suitable measures to carry out this evaluation (Schatzmann et al., 2005; Ai et al., 2007a; Ai et al., 2007b; Ai and Litman, 2006; Ai and Litman, 2007; Turunen et al., 2006).

In this paper, we use three different methodologies to acquire three different corpora for the same task. Our goal is to compare two corpora acquired using two different simulation techniques with a corpus acquired by real users. We use several evaluation measures proposed in previous research to evaluate how realistic and exploratory the simulated corpora are. Our study has been performed using a dialog system called *Facilisimo*. The domain of the system is the telephone access to a customer support service that helps solving simple and routine software/hardware repairing problems, both at the domestic and professional levels.

The remainder of the paper is organized as follows. Sections 2 reviews different approaches related to user modeling within the fields of language processing and conversational agents. Section 3 describes the proposed models for user simulation and automatic dialog generation. Section 4 describes the proposed statistical methodology for dialog management. Section 5 describes the practical application of our proposal for the *Facilisimo* conversational agent and the acquisition of a initial corpus for this task. Section 6 presents the process and the measures used to evaluate the quality of dialogs acquired with the different techniques. Section 7 shows the results of the comparison of the measures for the three corpora acquired for the *Facilisimo* system. Finally, some conclusions and future work lines are described in Section 8.

2. State of the art

The term computer simulation is defined in (Bandini et al., 2009) as "the usage of a computational model to gain additional insight into a complex system's behavior (e.g. biological or social systems) by envisioning the implications of the modeling choices, but also to evaluate designs and plans without actually bringing them into existence in the real world". Agent-based simulation (ABS) is a relatively recent modeling technique widely used to model these complex systems with applications in many disciplines ranging from logistics optimization (Weyns et al., 2006), biological systems (Bandini et al., 2006), traffic conditions (Balmer and Nagel, 2006), social sciences (Pavón et al., 2008), economics (Windrum et al., 2007), pedestrian simulation (A.L. Ballinas and Rangel, 2011), or urban planning (Navarro et al., 2011). Detailed studies can be found in (Macal and North, 2010; Bandini et al., 2009; Heath et al., 2009). The use of ABS models attempts to different causes, for instance, the system has still not fully completed, ethical reasons (e.g., the safety of humans would be involved), practical reasons (e.g., reduce the time and costs that are required to develop and evaluate the system), etc.

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 (Moller et al., 2006), 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 of simulated data should contribute positively to the development of the system.

Simulated data 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 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 of simulated data are extremely valuable for this purpose.

Two main approaches can be distinguish to the creation of simulated users: rule based and data or corpus based. In a rule-based simulated user the investigator can create different rules that determine the behavior of the system (Chung, 2004; Lin and Lee, 2001; Lopez-Cozar et al., 2003). This approach is particularly useful when the purpose of the investigation is to evaluate the effects of different dialog management strategies. In this way the investigator has complete control over the design of the evaluation study.

An alternative approach, often described as corpus-based or data-based, uses 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 (Schatzmann et al., 2006).

Research on data-driven approaches to dialog structure modeling is relatively new and focuses mainly on recognizing a structure of a dialog as it progresses. Dialog segmentation can be defined as the process of dividing up a dialog by one of several related notions (speaker's intention, topic flow, coherence structure, cohesive devices, etc.), identifying boundaries where the discourse changes taken into account such as specific criteria. This detection is usually based on combining different kinds of features, such as semantic similarities, inter-sentence similarities, entity repetition, word frequency, linguistic features, and prosodic and acoustic characteristics.

Different studies have been carried out for identifying discourse segments in spoken and written documents. For instance, TextTiling (Hearst, 1994) is a two step algorithm for the segmentation of texts into discourse units that are meant to reflect the topic flow on the text. Yamron (Yamron, 1998) presented an approach to segmentation that models an unbroken text stream as an unlabeled sequence of topics using Hidden Markov Models. Ponte presented in (Ponte and Croft, 1997) an approach based on information retrieval methods that map a query text into semantically related words and phrases. Passoneau and Litman developed an algorithm for identifying topic boundaries that uses decision trees to combine multiple linguistic features extracted from corpora of spoken text (Passoneau and Litman, 1997).

There is also a wide range of natural language processing applications for which discourse segmentation assists in. For instance, Angheluta, Busser and Moens adapted a three-step segmentation algorithm for automatic text summarization (Angheluta et al., 2002). Walker applies this kind of techniques for anaphora resolution (Walker, 1998). Different studies show the benefits of using discourse segmentation for question answering tasks (Chai and Jin, 2004).

In the literature, there are different studies for comparing human-agent (HA) and human-human (HH) interactions. Most of them compare specific features of both kind of conversations (Doran et al., 2001). In our work, we

try to infer the dialog structure of HH corpora by means of an active learning approach based on training an initial dialog model using the HA corpus and then use this model to learn the structure of HH conversations. To achieve this goal, two problems need to be addressed: i) creating a dialog representation that is suitable for representing the required domain-specific information, and ii) developing a machine learning approach that uses this representation to capture information from a corpus of in-domain conversations. This field presents as a main challenge the need of detecting the dialog segment using different data sources that are provided by both user and system entities during the course of the dialog (semantic information, confidence scores, task-dependent and independent information, etc.).

3. Proposed methodologies for user modeling

As it has been described in the introduction section, given the number of operations that must be carried out by a spoken dialog system, the scheme used for the development of these systems usually includes several generic modules that must cooperate to satisfy the user's requirements. The *Automatic Speech Recognition module* (ASR) transforms the user utterance into the most probable sequence of words. The *Spoken Language Understanding module* (SLU) provides a semantic representation of the meaning of the sequence of words generated by the ASR module. The *Dialog Manager* (DM) determines the next action to be taken by the system following a dialog strategy. The *Repository Query Manager* (RQM) receives requests for information or services, and returns the result to the dialog manager. The *Natural Language Generator module* (NLG) receives a formal representation of the system action and generates a user response in natural language. Finally, a *Text to Speech Synthesizer* (TTS) generates the audio signal transmitted to the user. The following subsections propose the use of two methodologies for modeling the user intention, which can be used to replace the ASR and SLU modules in the described architecture. The outputs generated by the user models are considered by the dialog manager to select the next system action by means of a statistical methodology that will be explained in Section 4.

3.1 First method for modeling the user intention

The first methodology that we have developed for modeling the user intention extends our previous work in statistical models for dialog management (Griol et al., 2014). Our proposed technique for user modeling simulates the user intention level by means of providing the next user dialog act in the same representation defined for the natural language understanding module. The lexical, syntactic and semantic information (e.g., words, part of speech tags, predicate-arguments structures, and named entities) associated to speaker u's ith clause is denoted as c_i^u.

Our model is based on the proposed in (Bangalore et al., 2008). In this model, each user clause is modeled as a realization of a user action defined by a subtask to which the clause contributes, the dialog act of the clause, and the named entities of the clause. For speaker u, DA_i^u denotes the dialog label of the ith clause, and ST_i^u denotes the subtask label to which the ith clause contributes. The dialog act of the clause is determined from the information about the clause and the previous dialog context (i.e., k previous utterances) as shown in Eq. 1.

$$DA_i^u = \underset{d^u \in \mathcal{DA}}{\operatorname{argmax}} P(d^u | c_i^u, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \quad (1)$$

In a second stage, the subtask of the clause is determined from the lexical information about the clause, the dialog act assigned to the clause according to Eq. 1, and the dialog context, as shown in Eq. 2.

$$ST_i^u = \underset{s^u \in \mathcal{ST}}{\operatorname{argmax}} P(s^u | DA_i^u, c_i^u, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \quad (2)$$

In our proposal, we consider both static and dynamic features to estimate the conditional distributions shown in Eq. 1 and 2. Dynamic features include the dialog act of each utterance and the task/subtask of each utterance. Static features include the words ans the part of speech tags in each utterance. As described in (Bangalore et al., 2008), the conditional distributions shown in Eq. 1 and 2 can be estimated by means of the general technique of choosing the maximum entropy (MaxEnt) distribution that properly estimates the average of each feature in the training data. This can be written as a Gibbs distribution parameterized with weights λ as Eq. 3 shows, where V is the size of the label set, X denotes the distribution of dialog acts or subtasks (DA_i^u or ST_i^u) and Φ denotes the vector of the described static and dynamic features used for the user turns from $i-1 \cdots i-k$.

$$P(X = st_i|\phi) = \frac{e^{\lambda_{st_i} \cdot \phi}}{\sum_{st=1}^{V} e^{\lambda_{st_i} \cdot \phi}} \tag{3}$$

Each of the classes can be encoded as a bit vector such that, in the vector corresponding to each class, the ith bit is one and all other bits are zero. Then, V-one-versus-other binary classifiers are used as Eq. 4 shows.

$$P(y|\phi) = 1 - P(\overline{y}|\phi) = \frac{e^{\lambda_y \cdot \phi}}{e^{\lambda_y \cdot \phi} + e^{\lambda_{\overline{y}} \cdot \phi}} = \frac{1}{1 + e^{-\lambda'_{\overline{y}} \cdot \phi}} \tag{4}$$

where $\lambda_{\overline{y}}$ is the parameter vector for the anti-label $\overline{y}$ and $\lambda'_{\overline{y}} = \lambda_y - \lambda_{\overline{y}}$.

3.2 Second method for modeling the user intention

The second method proposed for modeling the user intention is focused on the simulation of the user and conversational agents to acquire a dialog corpus. In our dialog generation technique, both agents use a random selection of one of the possible responses defined for the semantics of the task (expressed in terms of user and system dialog acts). At the beginning of the simulation, the set of system responses is defined as equiprobable. When a successful dialog is simulated, the probabilities of the answers selected by the the conversational agent simulator during that dialog are incremented before beginning a new simulation.

One of the main problems which must be considered during the interaction with a conversational agent is the propagation of errors through the different modules in the system. The recognition module must deal with the effects of spontaneous speech and with noisy environments; consequently, the sentence provided by this module could incorporate some errors. The understanding module could also add its own errors (which are mainly due to the lack of coverage of the semantic domain). Finally, the semantic representation provided to the dialog manager might also contain certain errors. Therefore, it is desirable to provide the dialog manager with information about what parts of the user utterance have been clearly recognized and understood and what parts have not.

In our proposal, the user simulator provides the conversational agent with the semantic representation associated to the user input together with its confidence scores (Garcia et al., 2003). To do this, an error simulation agent has been implemented to include semantic errors in the generation of dialogs. This agent modifies the dialog acts provided by the user agent simulator once it has selected the information to be provided to the user. In addition, the error simulation module adds a confidence score to each concept and attribute in the semantic representation generated for each user turn.

For the study presented in this paper, we have improved this agent using a model for introducing errors based on the method described in (Schatzmann et al., 2007). The generation of confidence scores is carried out separately from the model employed for error generation. This model is represented as a communication channel by means of a generative probabilistic model $P(c, a_u|\tilde{a}_u)$, where a_u is the true incoming user dialog act, $\tilde{a}_u$ is the recognized hypothesis, and c is the confidence score associated with this hypothesis.

The probability $P(\tilde{a}_u|a_u)$ is obtained by Maximum-Likelihood using the initial labeled corpus acquired with real users and considers the recognized sequence of words w_u and the actual sequence uttered by the user $\tilde{w}_u$. This probability is decomposed into a component that generates a word-level utterance from a given user dialog act, a model that simulates ASR confusions (learned from the reference transcriptions and the ASR outputs), and a component that models the semantic decoding process.

$$P(\tilde{a}_u|a_u) = \sum_{\tilde{w}_u} P(a_u|\tilde{w}_u) \sum_{w_u} P(\tilde{w}_u|w_u) P(w_u|a_u)$$

Confidence score generation is carried out by approximating $P(c|\tilde{a}_u, a_u)$ assuming that there are two distributions for c. These two distributions are handcrafted, generating confidence scores for correct and incorrect hypotheses by sampling from the distributions found in the training data corresponding to our initial corpus.

$$P(c|a_w, \tilde{a}_u) = \begin{cases} P_{corr}(c) & if \quad \tilde{a}_u = a_u \\ P_{incorr}(c) & if \quad \tilde{a}_u \neq a_u \end{cases}$$

The conversational agent simulator considers that the dialog is unsuccessful when one of the following conditions takes place:

- the dialog exceeds a maximum number of system turns empirically determined for each specific application domain;
- the response selected by the DM corresponds to a query not made by the user simulator;
- the RQM module generates an error because the user model has not provided the mandatory data needed to carry out the query;
- the NLG module generates an error when the response selected by the DM involves the use of a data item not provided by the user model.

A user request for closing the dialog is selected once the conversational agent simulator has provided the information defined in its objective(s). The dialogs that fulfill this condition before the maximum number of turns are considered successful.

4. Modeling the Dialog management process

In order to control the interactions with the user, our proposed statistical dialog management technique represents dialogs as a sequence of pairs (A_i, U_i), where A_i is the output of the dialog system (the system answer) at time i, and U_i is the semantic representation of the user turn (the result of the understanding process of the user input) at time i; both expressed in terms of dialog acts (Griol et al., 2014). This way, each dialog is represented by:

$$(A_1, U_1), \cdots, (A_i, U_i), \cdots, (A_n, U_n)$$

where A_1 is the greeting turn of the system, and U_n is the last user turn. We refer to a pair (A_i, U_i) as S_i, the state of the dialog sequence at time i.

In this framework, we consider that, at time i, the objective of the dialog manager is to find the best system answer A_i. This selection is a local process for each time i and takes into account the previous history of the dialog, that is to say, the sequence of states of the dialog preceding time i:

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i|S_1, \cdots, S_{i-1}) \tag{5}$$

where set $\mathcal{A}$ contains all the possible system answers.

Following Eq. 5, the dialog manager selects the following system prompt by taking into account the sequence of previous pairs (A_i, U_i). The main problem to resolve this equation is regarding the number of possible sequences of states, which is usually very large. To solve the problem, we define a data structure in order to establish a partition in this space (i.e., in the history of the dialog preceding time i). This data structure, which we call *Interaction Register* (IR), contains the sequence of user dialog acts provided by the user throughout the previous history of the dialog (i.e., the output of the NLU module).

After applying these considerations and establishing the equivalence relation in the histories of dialogs, the selection of the best A_i is given by Eq. 6.

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | IR_{i-1}, S_{i-1}) \tag{6}$$

We propose the use of a classification process to decide the next system action following the previous equation. From our previous work on dialog management (Griol et al., 2014), we propose the use of a multilayer perceptron for the classification, where the input layer receives the current state of the dialog, which is represented by the term (IR_{i-1},A_i). The values of the output layer can be viewed as the a posteriori probability of selecting the different user intention given the current situation of the dialog.

5. Case application: The Facilisimo spoken dialog system

We have applied our proposal to the problem solving domain of the *Facilisimo* spoken dialog system, which acts as a customer support service to help solving simple and routine software/hardware repairing problems, both at the domestic and professional levels. The system has been developed using an hybrid dialog management approach that combines the VoiceXML standard with the proposed user modeling and statistical dialog management techniques (Griol et al., 2012).

The definition of the system's functionalities and dialog strategy was carried out by means of the analysis of 150 human-human (HH) conversations provided by real assistants attending the calls of users with a software/hardware problem at the City Council of Leganés (Madrid, Spain). The labeling defined for this corpus contains different types of information, that have been annotated using a multilevel approach similar to the one proposed in the Luna Project (Stepanov et al., 2014). The first levels include segmentation of the corpus in dialog turns, transcription of the speech signal, and syntactic preprocessing with POS-tagging and shallow parsing. The next level consists of the annotation of main information using attribute-value pairs. The other levels of the annotation show contextual aspects of the semantic interpretation. These levels include the predicate structure, the relations between referring expressions, and the annotation of dialog acts.

The attribute-value annotation uses a predefined domain ontology to specify concepts and their relations. The attributes defined for the task include *Concept*, *Computer-Hardware*, *Action*, *Person-Name*, *Location*, *Code*, *TelephoneNumber*, *Problem*, etc.

Dialog act (DA) annotation was performed manually by one annotator on speech transcriptions previously segmented into turns. The DAs defined to label the corpus are the following: i) Core DAs: *Action-request*, *Yes-answer*, *No-answer*, *Answer*, *Offer*, *ReportOnAction*, *Inform*; ii) Conventional DAs: *Greet*, *Quit*, *Apology*, *Thank*; iii) Feedback-Turn management DAs: *ClarificationRequest*, *Ack*, *Filler*; iv) Non interpretable DAs: *Other*.

The original FrameNet[1] description of frames and frame elements was adopted for the predicate-argument structure annotation, introducing new frames and roles related to hardware/software only in case of gaps in

[1] https://framenet.icsi.berkeley.edu/fndrupal/

the FrameNet ontology. Some of the frames included in this representation are *Telling*, *Greeting*, *Contacting*, *Statement*, *Recording*, *Communication*, *Being operational*, *Change operational state*, etc.

An example of the attribute-value, dialog-act and predicate structure annotations of a user utterance is shown below:

Hi, I have a problem with my printer.
Attributes-values: *Concept*:problem; *Hardware*:printer;
Dialog acts: *Answer*;
Predicate structure: (*Greeting*)(*Problem_description*) *Device Problem*

The basic structure of the dialogs is usually composed by the sequence of the following tasks: *Opening*, *Problem-statement*, *User-identification*, *Problem-clarification*, *Problem-resolution*, and *Closing*. This set of tasks contains a list of subtasks, such as *Problem-description*, *Problem-Request*, *Problem-Confirmation*, *Brand-Identification*, *Model-Identification*, *Help-Request*, *Message-Confirmation*, *Name-Identification*, etc. The shared plan is represented as a data register that encapsulates the task structure, dialog act structure, attribute-values and predicate-argument structure of utterances.

During the *Problem-statement* task, the caller explains the problem the reasons why he/she calls the help-desk. In the *User-identification* task, the operator asks for additional information regarding the identity of the caller. Once the caller has described the problem, the operator can ask for additional information to clarify it during the *Problem-clarification* task.

During the *Problem-resolution* task, the operator asks the user to perform specific tests. We have defined nine different subtasks inside this generic segment, given that our goal is to detect not only that the dialog is in this segment, but also what are the specific problem that has to be resolved: *Printer* (P4), *Network connection* (P5), *PC going slow* (P6), *Monitor* (P7), *Keyboard* (P8), *Mouse* (P9), *CD-DVD player* (P10), *Power supply* (P11), and *Virus* (P12). The operator assigns a ticket number for the current call if the problem has not been solved after this task. The user must take note of this number and inform about this to the operator. The dialog ends at the *Closing* phase, in which the operator also tries to give a useful advice related to the described problem.

A total of 150 dialogs was acquired by 25 recruited users using the developed system (Human-Agent dialogs, HA). Ten different dialog scenarios inspired from the services provided in the human-human dialogs were designed for the acquisition. The complete set of HH and HA dialogs were manually labeled including this task/subtask information. This information was incorporated for each user and system turn in the dialogs. Figure 1 shows the transcription of a dialog included in the HA corpus.

6. Measures defined for the Evaluation

For the evaluation of the quality of the dialogs provided by a conversational agent, we have defined a set of quantitative evaluation measures based on prior work in the dialog literature (Schatzmann et al., 2005; Ai et al., 2007a). This set of proposed measures can be divided into two types:

- High-level dialog features: These features evaluate how long the dialogs last, how much information is transmitted in individual turns, and how active the dialog participants are.
- Dialog style/cooperativeness measures: These measures analyze the frequency of different speech acts and study what proportion of actions is goal-directed, what part is taken up by dialog formalities, etc.

Six high-level dialog features have been defined for the evaluation of the dialogs: the average number of turns per dialog, the percentage of different dialogs without considering the attribute values, the number of

S1: Welcome to the Facilisimo system. How may I help you?
U1: Hello, my printer does not print correctly.
S2: Tell me your name and surname.
U2: My name is Ricardo Santamarta.
S3: Can you tell me your phone extension?
U3: Nine four sixty-three.
S4: Can you tell me the brand and model?
U4: HP Officejet Pro 8610.
S5: Is the printer on now?
U5: It is on.
S6: Can you tell me the message on the printer display?
U6: Paper jam is written.
S7: Can you see the active printer from the control panel of your PC?
U7: It is active.
S8: Please try the following steps. Say next after each step. Turn the printer off, and then on again.
U8: Next.
S9: Print a Printer Status Report.
U9: Next.
S10: Has this process solved your problem?
U10: No.
S11: Remove any loose paper from the main and optional trays.
U11: Next.
S12: Clear any jammed paper from the rear of the printer.
U12: Next.
S13: Has this process solved your problem?
U13: Yes.
S14: Anything else?
U14: No, thank you.
S15: Most of these steps will also work for troubleshooting an office copier. Have a nice day!

Figure 1: Example of a dialog acquired with recruited users and the Facilisimo dialog system (translation from Spanish to English)

repetitions of the most seen dialog, the number of turns of the most seen dialog, the number of turns of the shortest dialog, and the number of turns of the longest dialog. Using these measures, we tried to evaluate the success of the simulated dialogs as well as its efficiency and variability of the dialogs.

For dialog style features, we have defined and counted a set of system/user dialog acts. On the system side, we have measured the confirmation of concepts and attributes, questions to require information, and system answers generated after a database query. On the user side, we have measured the percentage of turns in which the user carries out a request to the system, provide information, confirms a concept or attribute, Yes/No answers, and other answers not included in the previous categories.

7. Evaluation Results

To compare the different dialog corpus, we compute the mean value for each corpus with respect to each of the evaluation measures shown in the previous section. A set of 3000 successful dialogs was simulated with each one of the simulation techniques using the same scenarios defined for the acquisition with real users. Two-tailed t-tests have been used to compare the means across the three corpora as described in (Ai et al., 2007a). All differences reported as statistically significant have p-values less than 0.05 after Bonferroni corrections. The following notation has been introduced to present the results of the evaluation of the different measures: **Corpus 1** (corpus acquired with real users), **Corpus 2** (corpus acquired by means of the statistical user simulation technique), and **Corpus 3** (corpus acquired by means of the automatic dialog generation technique).

7.1 High-level Dialog Features

As stated in the previous section, the first group of experiments covers the following statistical properties: i) Dialog length, measured in the average number of turns per dialog, number of turns of the shortest dialog, number of turns of the longest dialog, and number of turns of the most seen dialog; ii) Different dialogs in each corpus, measured in the percentage of different dialogs and the number of repetitions of the most seen dialog; iii) Turn length, measured in the number of actions per turn; iv) Participant activity as a ratio of system and user actions per dialog.

Table 1 shows the results of the comparison of the high-level dialog features. It can be seen that all measures have similar values in the three corpora. The more significant difference is the average number of user turns. In the two types of scenarios, the dialogs acquired using the simulation technique are shorter than those acquired with real users. This fact can be explained due there are a set of dialogs acquired with real users in which the user asked for additional information not included in the definition of the corresponding scenario once its objectives were achieved.

	Corpus 1	**Corpus 2**	**Corpus 3**
Average number of user turns per dialog	11.99	10.75	11.17
Percentage of different dialogs	84.73%	79.45%	86.32%
Number of repetitions of the most seen dialog	5	11	4
Number of turns of the most seen dialog	9	7	7
Number of turns of the shortest dialog	5	5	5
Number of turns of the longest dialog	16	14	16

Table 1: Results of the high-level dialog features defined for the comparison of the three corpora

The mean values of the turn length and dialog length for the real and statistical corpus are almost the same. The dialogs acquired with the automatic generation technique are statistically shorter, as they provide 1.17 actions per user turn instead of the 1.06 actions provided by the real users and the corpus acquired by means of the statistical dialog simulation technique. The shape of the distributions shows that the real dialogs have the largest standard deviation given that the task length of these dialogs is more disperse. The dialogs acquired using the automatic generation technique have the minor deviation since the successful dialogs are usually those that use the minor number of turns to achieve the objective(s) predefined. Regarding the dialog participant activity, the real and statistical corpora have almost exact values for the ratio of user versus system actions. The proportion of system actions in the corpus acquired using the automatic dialog generation technique is only slightly higher.

7.2 Dialog style and cooperativeness

Finally, we have compared the percentage of the most significant types of dialog acts in both corpora. Tables 2 and 3 respectively show the frequency of the most dominant user and system dialog acts. It can be observed that the relative ordering of user actions is similar for the three corpora. The most significant difference between the three corpora is the percentage of turns in which the user provides information. The percentage of this kind of answer is higher in the real corpus. This can be explained by the fact that it is less probable that simulated users provide additional information not included as mandatory for the corresponding scenario. For this reason, there is a higher percentage of system answers that inform the user in the real corpus.

	Corpus 1	**Corpus 2**	**Corpus 3**
Confirmation of concepts and attributes	13.76%	12.63%	19.43%
Questions to require information	38.24%	36.97%	37.01%
Answers generated after a database query	48.00%	50.40%	43.56%

Table 2: Percentages of the different types of system dialog acts in the three corpora

	Corpus 1	**Corpus 2**	**Corpus 3**
Request to the system	31.04%	34.23%	33.92%
Provide information	21.62%	20.87%	21.13%
Confirmation	10.21%	9.34%	8.09%
Yes/No answers	35.02%	34.10%	36.21%
Other answers	2.11%	1.46%	0.65%

Table 3: Percentages of the different types of user dialog acts in the three corpora

8. Conclusions

In this paper, we investigated the differences between three dialog corpus acquired using different techniques. A statistical user simulation technique and an automatic dialog generation technique have been developed in order to acquire simulated corpus to be compared with a previous corpus acquired by real users. We have used a set of well-known dialog evaluation measures to evaluate statistically the differences between the different corpora.

The results of the evaluation of this methodology in a practical problem solving task show that, although there are some perceptible differences in the frequency of particular system/user dialog acts, there are no significant differences in the overall values obtained for the different measures. The development of two methodologies that simulates the user intention level makes also possible the training of dialog models which a very similar behavior, as we have studied in previous works.

We are adapting the proposed simulations techniques to acquire different dialog corpus within the framework of more complicated domains, as well as the study of new techniques to carry out error simulation and consider new information sources required to model the user's emotional state and personality traits.

9. Acknowledgements

This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

10. References

Ai, H. and Litman, D., 2006. Comparing Real-Real, Simulated-Simulated, and Simulated-Real Spoken Dialogue Corpora. In *Procs. of AAAI Workshop Statistical and Empirical Approaches for Spoken Dialogue Systems*. Boston, USA.

Ai, H. and Litman, D., 2007. Knowledge Consistent User Simulations for Dialog Systems. In *Proc. of Interspeech'07*, pages 2697–2700. Antwerp, Belgium.

Ai, H., Raux, A., Bohus, D., Eskenazi, M., and Litman, D., 2007a. Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 124–131. Antwerp, Belgium.

Ai, H., Tetreault, J., and Litman, D., 2007b. Comparing User Simulation Models For Dialog Strategy Learning. In *Proc. of NAACL HLT'07*, pages 1–4. Rochester, NY, USA.

A.L. Ballinas, A. M. and Rangel, A., 2011. Multiagent System Applied to the Modeling and Simulation of Pedestrian Traffic in Counterflow. *Journal of Artificial Societies and Social Simulation*, 14(3).

Angheluta, R., Busser, R. D., and Moens, M., 2002. The use of topic segmentation for automatic summarization. In *Proc. ACL Workshop on Automatic Summarization*, pages 66–70.

Bailly, G., Raidt, S., and Elisei, F., 2010. Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52(6):598–612.

Balmer, M. and Nagel, K., 2006. *Innovations in Design & Decision Support Systems in Architecture and Urban Planning*, chapter Shape Morphing of Intersection Layouts Using Curb Side Oriented Driver Simulation, pages 167–183. Springer-Verlag.

Bandini, S., Celada, F., Manzoni, S., Puzone, R., and Vizzari, G., 2006. Modelling the Immune System with Situated Agents. *Lecture Notes in Computer Science*, 3931:231–243.

Bandini, S., Manzoni, S., and Vizzari, G., 2009. Agent Based Modeling and Simulation: An Informatics Perspective. *Journal of Artificial Societies and Social Simulation*, 12(4):97–126.

Bangalore, S., Fabbrizio, G. D., and Stent, A., 2008. Learning the Structure of Task-driven Human-Human Dialogs. *IEEE Trans Audio Speech Lang Processing*, 16(7):1249–1259.

Bohus, D., Grau, S., Huggins-Daines, D., Keri, V., Krishna, G., Kumar, R., Raux, A., and Tomko, S., 2007. Conquest - An Open-Source Dialog System for Conferences. In *Proc. of 7th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL'07)*, pages 9–12. Rochester, USA.

Bohus, D. and Rudnicky, A., 2002. LARRI: A Language-Based Maintenance and Repair Assistant. In *Proc. of Multi-Modal Dialogue in Mobile Environments Conference (IDS'02)*. Kloster Irsee, Germany.

Brahnam, S., 2009. Building Character for Artificial Conversational Agents: Ethos, Ethics, Believability, and Credibility. *PsychNology Journal*, 7(1):9–47.

Chai, J. and Jin, R., 2004. Discourse structure for context question answering. In *Proc. HLT-NAACL Workshop on Pragmatics of Question Answering*, pages 23–30.

Chung, G., 2004. Developing a flexible spoken dialog system using simulation. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 63–70. Barcelona, Spain.

Doran, C., Aberdeen, J., Damianos, L., and Hirschman., L., 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proc. SigDial*.

Garcia, F., Hurtado, L., E.Sanchis, and Segarra, E., 2003. The incorporation of Confidence Measures to Language Understanding. In *Proc. of TSD'03*, pages 165–172. Ceske Budejovice.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., and Zue, V., 1995. Multilingual spoken-language understanding in the MIT Voyager system. *Speech Communication*, 17:1–18.

Griol, D., Callejas, Z., López-Cózar, R., and Riccardi, G., 2014. A domain-independent statistical methodology for dialog management in spoken dialog systems. *Computer, Speech and Language*, 28(3):743–768.

Griol, D., Molina, J., and Callejas, Z., 2012. Bringing together commercial and academic perspectives for the development of intelligent AmI interfaces. *JAISE*, 4(3):183–207.

Hearst, M., 1994. Multi-paragraph segmentation of expository text. In *Proc. ACL*, pages 9–16.

Heath, B., Hill, R., and Ciarallo, F., 2009. A Survey of Agent-Based Modeling Practices (January 1998 to July 2008). *Journal of Artificial Societies and Social Simulation*, 12(4).

Heinroth, T. and Minker, W., 2012. *Introducing Spoken Dialogue Systems into Intelligent Environments*. Kluwer Academic PublishersSpringer-Verlag.

Lin, B. and 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.

Lopez-Cozar, R., la Torre, A. D., Segura, J., Rubio, A., and S□nchez, V., 2003. Assessment of dialogue systems by means of a new simulation technique. *Speech Communication*, 40(3):387–407.

Macal, C. and North, M., 2010. Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4:151–162.

Melin, H., Sandell, A., and Ihse, M., 2001. CTT-bank: A speech controlled telephone banking system - an initial evaluation. In *TMH Quarterly Progress and Status Report (TMH-QPSR)*, volume 1, pages 1–27.

Menezes, P., Lerasle, F., Dias, J., and Germa, T., 2007. *Humanoid Robots, Human-like Machines*, chapter Towards an Interactive Humanoid Companion with Visual Tracking Modalities, pages 48–78. Advanced Robotic Systems Int. and I-Tech Education and Publishing.

Moller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., and Reithinger, N., 2006. MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In *Proc. of the 9th Int. Conference on Spoken Language Processing (Interspeech/ICSLP)*, pages 1786–1789. Pittsburgh, USA.

Navarro, L., Flacher, F., and Corruble, V., 2011. Dynamic Level of Detail for Large Scale Agent-Based Urban Simulations. In *Proc. of the 10th Int. Conference on Autonomous agents and multiagent systems (AAMAS'11)*, pages 701–708. Taipei, Taiwan.

Passoneau, R. and Litman, D., 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23:103–139.

Pavón, J., Sansores, C., Gómez, J., and Wang, F., 2008. Modelling and simulation of social systems with INGENIAS. *Int. Journal of Agent-Oriented Software Engineering*, 2(2).

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

Pieraccini, R., 2012. *The Voice in the Machine: Building Computers that Understand Speech*. The MIT Press.

Ponte, J. and Croft, W., 1997. Text segmentation by topic. In *Proc. ECDL*, pages 120–129.

Schatzmann, J., Georgila, K., and Young, S., 2005. Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. In *Proc. of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 45–54. Lisbon, Portugal.

Schatzmann, J., Thomson, B., and Young, S., 2007. Error Simulation for Training Statistical Dialogue Systems. In *Proc. of ASRU'07*, pages 526–531.

Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S., 2006. A Survey of Statistical User Simulation

Techniques for Reinforcement-Learning of Dialogue Management Strategies. *Knowledge Engineering Review*, 21(2):97–126.

Stepanov, E., Riccardi, G., and Bayer, A., 2014. The Development of the Multilingual LUNA Corpus for Spoken Language System Porting. In *Proc. LREC*, pages 2675–2678.

Turunen, M., Hakulinen, J., and Kainulainen, A., 2006. Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences. In *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP)*, pages 1057–1060. Pittsburgh, USA.

Vaquero, C., Saz, O., Lleida, E., Marcos, J., and Canal□s, C., 2006. VOCALIZA: An application for computer-aided speech therapy in spanish language. In *Proc. IV Jornadas en Tecnolog□a del Habla*, pages 321–326. Zaragoza, Spain.

Walker, M. A., 1998. *Centering, anaphora resolution, and discourse structure*, pages 401–435. Oxford University Press.

Weng, F., Varges, S., Raghunathan, B., Ratiu, F., Pon-Barry, H., Lathrop, B., Zhang, Q., Scheideck, T., Bratt, H., Xu, K., Purver, M., Mishra, R., Raya, M., Peters, S., Meng, Y., Cavedon, L., and Shriberg, L., 2006. CHAT: A Conversational Helper for Automotive Tasks. In *Proc. of the 9th Int. Conference on Spoken Language Processing (Interspeech/ICSLP)*, pages 1061–1064. Pittsburgh, USA.

Weyns, D., Boucké, N., and Holvoet, T., 2006. Gradient Field-Based Task Assignment in an AGV Transportation System. In *Proc. of the 5th Int. Conference on Autonomous agents and multiagent systems (AAMAS'06)*, pages 842–849. Hakodate, Japan.

Windrum, P., Fagiolo, G., and Moneta, A., 2007. Empirical Validation of Agent-Based Models: Alternatives and Prospects. *Journal of Artificial Societies and Social Simulation*, 10(2).

Yamron, J., 1998. Topic detection and tracking segmentation task. In *Proc. Broadcast News Transcription and Understanding Workshop*.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., and Hetherington, L., 2000. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96.