Predicting the Mental State of the Users in Spoken Dialogue Systems

Zoraida Callejas[1], David Griol[2], Ramón López-Cózar[1]

zoraida@ugr.es, dgriol@inf.uc3m.es, rlopezc@ugr.es

1 Dept. of Languages and Computer Systems, CITIC-UGR. University of Granada. C/ Pdta. Daniel Saucedo Aranda, 18071, Granada (Spain)

2 Dept. of Computer Science, Carlos III University of Madrid, Av. Universidad, 30, 28911, Leganés (Spain)

Abstract

In this paper we propose a method for predicting the user mental state for the development of more efficient and usable spoken dialogue systems. This prediction, carried out for each user turn in the dialogue, makes it possible to adapt the system dynamically to the user needs. The mental state is built on the basis of the emotional state of the user and their intention, and is recognized by means of a module conceived as an intermediate phase between natural language understanding and the dialogue management in the architecture of the systems. We have implemented the method in the UAH system, for which the evaluation results with both simulated and real users show that taking into account the user's mental state improves system performance as well as its perceived quality.

1. Introduction

In human conversation, speakers adapt their message and the way they convey it to their interlocutors and to the context in which the dialogue takes place. Thus, the interest in developing systems capable of maintaining a conversation as natural and rich as a human conversation has fostered research on adaptation of these systems to the users.

For example, Jokinen (2003) describes different levels of adaptation. The simplest one is through personal profiles in which the users make static choices to customize the interaction (e.g. whether they want a male or female system's voice), which can be further improved by classifying users into preferences' groups. Systems can also adapt to the user environment, as in the case of Ambient Intelligence applications (Ábalos et al., 2010). A more sophisticated approach is to adapt the system to the user specific knowledge and expertise, in which case the main research topics are the adaptation of systems to proficiency in the interaction language (Ohkawa et al., 2009), age (Wolters et al., 2009), different user expertise levels (Evanini et al., 2008), and special needs (Miesenberger et al., 2010). Despite their complexity, these characteristics are to some extent rather static. Jokinen (2003) identifies a more complex degree of adaptation in which the system adapts to the user's intentions and state.

Most spoken dialogue systems that employ user *mental states* address these states as intentions, plans or goals. One of the first models of mental states was introduced by Ginzburg (1966) in his information state theory for dialogue management. According to this theory, dialogue is characterized as a set of actions in order to change the interlocutor's mental state and reach the goals of the interaction. This way, the mental state is addressed as the user's beliefs and intentions. During the last decades, this theory has been successfully applied to build spoken dialogue systems with a reasonable flexibility (Jokinen and McTear, 2010).

Another pioneer work which implemented the concept of mental state was the spoken dialogue system TRAINS-92 (Traum, 1993). This system integrated a domain plan reasoner which recognized the user mental state and used it as a basis for utterance understanding and dialogue management. The mental state was conceived as a dialogue plan which included goals, actions to be achieved and constraints in the plan execution.

More recently, some authors have considered mental states as equivalent to emotional states (Nisimura et al., 2006), given that affect is an evolutionary mechanism that plays a fundamental role in human interaction in order to adapt to the environment and carry out meaningful decision making (Callejas et al., 2011). As stated by Sobol-Shikler (2011), the term *affective state* may refer to emotions, attitudes, beliefs, intents, desires, pretending, knowledge, and moods.

Although emotion is gaining increasing attention from the dialogue systems community, most research described in the literature is devoted exclusively to emotion recognition. For example, a comprehensive and updated review can be found in (Schuller et al., 2011). In this paper we propose a mental state prediction method which takes into account both the users' intentions and their emotions, and describes how to incorporate such a state into the architecture of a spoken dialogue system to adapt dialogue management accordingly.

The rest of the paper is organized as follows. In Section 2 we describe the motivation of our proposal and related work. Section 3 presents in detail the proposed model and how it can be included into the architecture of a spoken dialogue system. To test the suitability of the proposal we have carried out experiments with the UAH system, which is described in Section 4 together with the annotation of a corpus of user interactions. Section 5 describes the methodology used to evaluate the proposal, whereas in Section 6 we discuss the evaluation results obtained by comparing the initial UAH system with an enhanced version of if that adapts its behaviour to the perceived user mental state. Finally, in Section 7 we present the conclusions and outline guidelines for future work.

2. Background

In traditional computational models of the human mind, it is assumed that mental processes respect the semantics of mental states, and the only computational explanation for such mental processes is a computing mechanism that manipulates symbols related to the semantic properties of mental states (Piccinini, 2004). However, there is no universally agreed-upon description of such semantics, and mental states are defined in different ways, usually ad hoc, even when they are shared as a matter of study in different disciplines.

Initially, mental states were reduced to a representation of the information that an agent or system holds internally and it uses to solve tasks. Following this approach, Katoh et al. (1998) proposed to use mental states as a basis to decide whether an agent should participate in an assignment according to its self-perceived proficiency in solving it. Using this approach, negotiation and work load distribution can be optimized in multi-agent systems. As they themselves claim, the authors' approach has no basis on the communication theory. Rather, the mental state stores and prioritizes features which are used for action selection. However, in spoken dialogue systems it is necessary to establish the relationship between mental states and the communicative acts.

Beun (1994) claimed that in human dialogue, speech acts are intentionally performed to influence "the relevant aspects of the mental state of a recipient". The author considers that a mental state involves beliefs, intentions and expectations. Dragoni (2008) followed this vision to formalize the consequences of an utterance or series of dialogue acts on the mental state of the hearer in a multi-context framework. This framework lied on a representation of mental states which coped only with beliefs (representations of the real state of the world) and desires (representations of an "ideal" state of the world). Other aspects which could be considered as mental states, such as intentions, had to be derived from these primitive ones.

The transitions between mental states and the situations that trigger them have been studied from other perspectives different from dialogue. For example, Jonker and Treur (2002) proposed a formalism for mental states and their properties by describing their semantics in temporal traces, thus accounting for their dynamic changes during interactions. However, they only considered physical values such as hunger, pain or temperature.

In psychophysiology, these transitions have been addressed by directly measuring the state of the brain. For example, Fairclough (2009) surveyed the field of psychophysiological characterization of the user states, and defined mental states as a representation of the progress within a task-space or problem-space. Das et al. (2009) presented a study on mental state estimation for Brain-Computer Interfaces, where the focus was on mental states obtained from the electrocorticograms of patients with medically intractable epilepsy. In this study, mental states were defined as a set of stages which the brain undergoes when a subject is engaged in certain tasks, and brain activity was the only way for the patients to communicate due to motor disabilities.

Other authors have reported dynamic actions and also physical movements as a main source of information to recognize mental states. For example, Sindlar et al. (2010) used dynamic logic to model ascription of beliefs, goals, or plans on grounds of observed actions to interpret other agents' actions. Oztop et al. (2005) developed a computational model of mental state inference that used the circuitry that underlied motor control. This way, the mental state of an agent could be described as the goal of the movement or the intention of the agent performing such movement. Lourens et al. (2010) also carried out mental state recognition from motor movements following the mirror neuron system perspective.

In the research described so far, affective information is not explicitly considered although it can sometimes be represented using a number of formalisms. However, recent work has highlighted the affective and social nature of mental states. This is the case of recent psychological studies in which mental states do not cope with beliefs, intentions or actions, but rather are considered emotional states. For example, Dyer et al. (2000) presented a study on the cognitive development of mental state understanding of children in which they discovered the positive effect of storybook reading to make children more effective being aware of mental states. The authors related English terms found in story books to mental states, not only using terms such as *think*, *know* or *want*, but also words that refer to emotion, desire, moral evaluation and obligation.

Similarly, Lee et al. (2005) investigated mental state decoding abilities in depressed women and found that they were significantly less accurate than non-depressed in identifying mental states from pictures of eyes. They accounted for mental states as beliefs, intentions and specially emotions, highlighting their relevance to understand behaviour. The authors also pointed out that the inability to decode and reason about mental states has a severe impact on socialization of patients with schizophrenia, autism, psychopathy and depression.

In (Osatuke and Stiles, 2010), the authors investigate the impairment derived from the inability to recognize others’ mental states as well as the impaired accessibility of certain self-states. This way, they involve into the concept of mental-state terms not only related to emotion (happy, sad, and fearful) but also to personality, such as assertive, confident or shy.

Sobol-Shikler (2011) shares this vision and proposes a representation method that comprises a set of affective-state groups or archetypes that often appear in everyday life. His method is designed to infer combinations of affective states that can occur simultaneously and whose level of expression can change over time within a dialogue. By affective states, the author understands moods, emotions and mental states. Although he does not provide any definition of mental state, the categories employed in his experiments do not account for intentional information.

In the area of dialogue systems, emotion has been used for several purposes, as summarized in the taxonomy of applications proposed by Batliner et al. (2006). In some application domains, it is fundamental to recognize the affective state of the user to adapt the systems behaviour. For example, in emergency services (Bickmore and Giorgino, 2004) or intelligent tutors (Litman and Forbes-Riley, 2006), 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 dialogue that cause negative emotional states, avoid them and foster positive ones in future interactions

Emotions affect the explicit message conveyed during the interaction. They change people’s voices, facial expressions, gestures, and speech speed; a phenomenon addressed as *emotional colouring* (Khalifa et al., 2007; Acosta and Ward, 2009). This effect can be of great importance for the interpretation of user input, for example, to overcome the Lombard effect in the case of angry or stressed users (Boril and Hansen, 2010), and to disambiguate the meaning of the user utterances depending on their emotional status (Bosma and Andre, 2004).

Emotions can also affect the actions that the user chooses to communicate with the system. According to Wilks et al. (2011), emotion can be understood more widely as a manipulation of the range of interaction affordances available to each counterpart in a conversation. Riccardi and Hakkani-Tür (2005) studied the impact of emotion temporal patterns in user transcriptions, semantic and dialogue annotations of the *How May I help you?* system. In their study, the representation of the user state was defined "only in terms of dialogue act or expected user intent". They found that emotional information can be useful to improve the dialogue strategies and predict system errors, but it was not employed in their system to adapt dialogue management.

Boril et al. (2010) measured speech production variations during the interactions of drivers with commercial automated dialogue systems. They discuss that cognitive load and emotional states affect the number of query repetitions required for the users to obtain the information they are looking for.

Baker et al. (2009) described a specific experience for the case of computer-based learning systems. They found that boredom significantly increases the chance that a student will game the system on the next observation. However, the authors do not describe any method to couple emotion and the space of afforded possible actions.

Gnjatovic and Rösner (2008) implemented an adapted strategy for providing support to users depending on their emotional state while they solved the Tower-of-Hanoi puzzle in the NIMITEK system. Although the help policy was adapted to emotion, the rest of the decisions of the dialogue manager were carried out without taking into account any emotional information.

In our proposal, we merge the traditional view of the dialogue act theory in which communicative acts are defined as intentions or goals, with the recent trends that consider emotion as a vital part of mental states that makes it possible to carry out social communication. To do so, we propose a mental state prediction module which can be easily integrated in the architecture of a spoken dialogue system and that is comprised of an intention recognizer and an emotion recognizer as explained in Section 3.

Delaborde and Devillers (2010) proposed a similar idea to analyze the immediate expression of emotion of a child playing with an affective robot. The robot reacted according to the prediction of the children emotional response. Although there was no explicit reference to “mental state”, their approach processed the child state and employed both emotion and the action that he would prefer according to an interaction profile. There was no dialogue between the children and the robot, as the user input was based mainly in non-speech cues. Thus, the actions that were considered in the representation of the children state are not directly comparable to the dialogue acts that we address in the paper.

Very recently, other authors have developed affective dialogue models which take into account both emotions and dialogue acts. The dialogue model proposed by (Pitterman et al. 2010) combined three different submodels: an emotional model describing the transitions between user emotional states during the interaction regardless of the data content, a plain dialogue model describing the transitions between existing dialogue states regardless of the emotions, and a combined model including the dependencies between combined dialogue and emotional states. Then, the next dialogue state was derived from a combination of the plain dialogue model and the combined model. The dialogue manager was written in Java embedded in a standard VoiceXML application enhanced with ECMAScript. In our proposal, we employ statistical techniques for inferring user acts, which makes it easier porting it to different application domains. Also the proposed architecture is modular and thus makes it possible to employ different emotion and intention recognizers, as the intention recognizer is not linked to the dialogue manager as in the case of Pitterman et al (2010).

Bui et al. (2009) based their model on Partially Observable Markov Decision Processes (Williams and Young, 2007) that adapt the dialogue strategy to the user actions and emotional states, which are the output of an emotion recognition module. Their model was tested in the development of a route navigation system for rescues in an unsafe tunnel in which users could experience five levels of stress. In order to reduce the computational cost required for solving the POMDP problem for dialogue systems in which many emotions and dialogue acts might be considered, the authors employ decision networks to complement POMDP. We propose an alternative to this statistical modeling which can also be used in realistic dialogue systems and evaluate it in a less emotional application domain in which emotions are produced more subtly.

3. New model for predicting the user mental state

We propose a model for predicting the user mental state which can be integrated in the architecture of a spoken dialogue system as shown in Figure 1. As can be observed, the model is placed between the natural language understanding (NLU) and the dialogue management phases. The model is comprised of an emotion recognizer, an intention recognizer and a mental state composer. The emotion recognizer detects the user emotional state by extracting an emotion category from the voice signal and the dialogue history. The intention recognizer takes the semantic representation of the user input and predicts the next user action. Then, in the mental state composition phase, a mental state data structure is built from the emotion and intention recognized and passed on to the dialogue manager.

An alternative to the proposed method would be to directly estimate the mental state from the voice signal, the dialogue features and the semantics of the user input in a single step. However, we have considered several phases that differentiate the emotion and intentions recognizers to provide a more modular architecture, in which different emotion and intention recognizers could be plugged-in. Nevertheless, we consider interesting as a future work guideline to compare this alternative estimation method with our proposal and check whether the performance gets improved, and if so, how to balance it with the benefits of modularization.

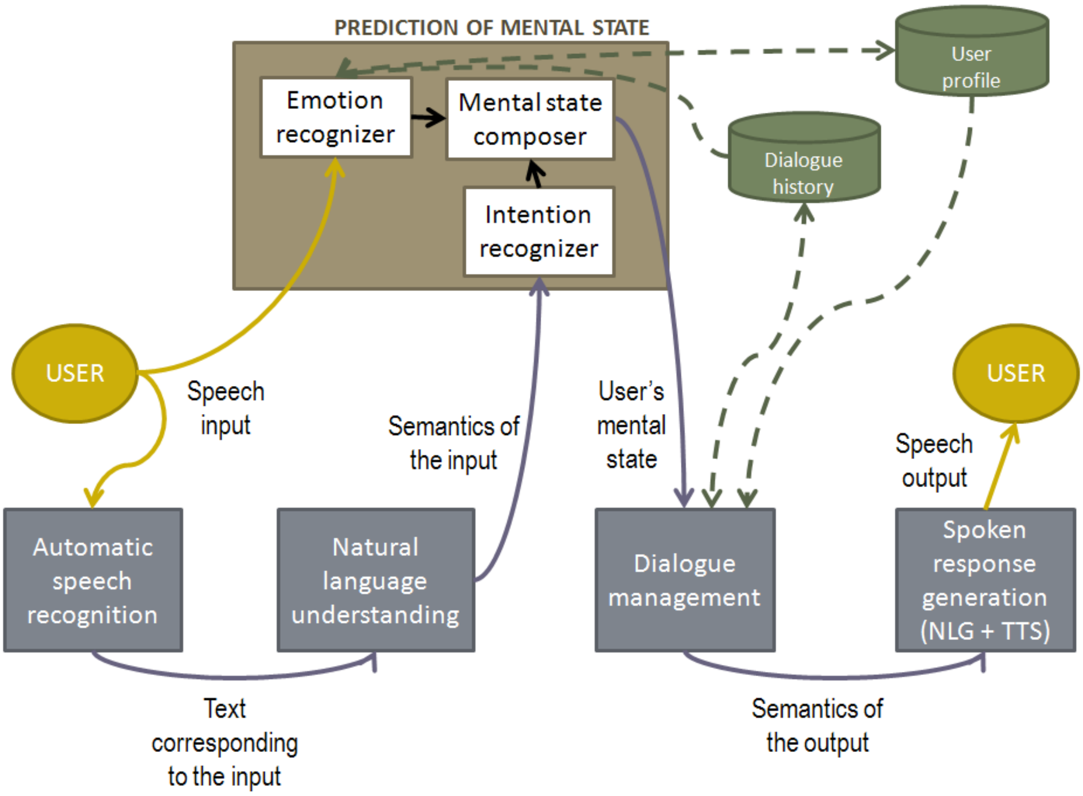


Figure 1: Integration of mental state prediction into the architecture of a spoken dialogue system

3.1. The emotion recognizer

As the architecture shown in Figure 1 has been designed to be highly modular, different emotion recognizers could be employed within it. We propose to use an emotion recognizer based solely in acoustic and dialogue information because in most application domains the user utterances are not long enough for the linguistic parameters to be significant for the detection of emotions. However, emotion recognizers which make use of linguistic information such as the one in (López-Cózar et al., 2008) can be easily employed within the proposed architecture by accepting an extra input with the result of the automatic speech recognizer.

Our recognition method, based on the previous work described in (Callejas and López-Cózar, 2008a), firstly takes acoustic information into account to distinguish between the emotions which are acoustically more different, and secondly dialogue information to disambiguate between those that are more similar.

We are interested in recognizing negative emotions that might discourage users from employing the system again or even lead them to abort an ongoing dialogue. Concretely, we have considered three negative emotions: *anger*, *boredom* and *doubtfulness*, where the latter refers to a situation in which the user uncertain about what to do next).

Following the proposed approach, our emotion recognizer employs acoustic information to distinguish *anger* from *doubtfulness* or *boredom* and dialogue information to discriminate between *doubtfulness* and *boredom*, which are more difficult to discriminate only by using phonetic cues. This process is shown in Figure 2.

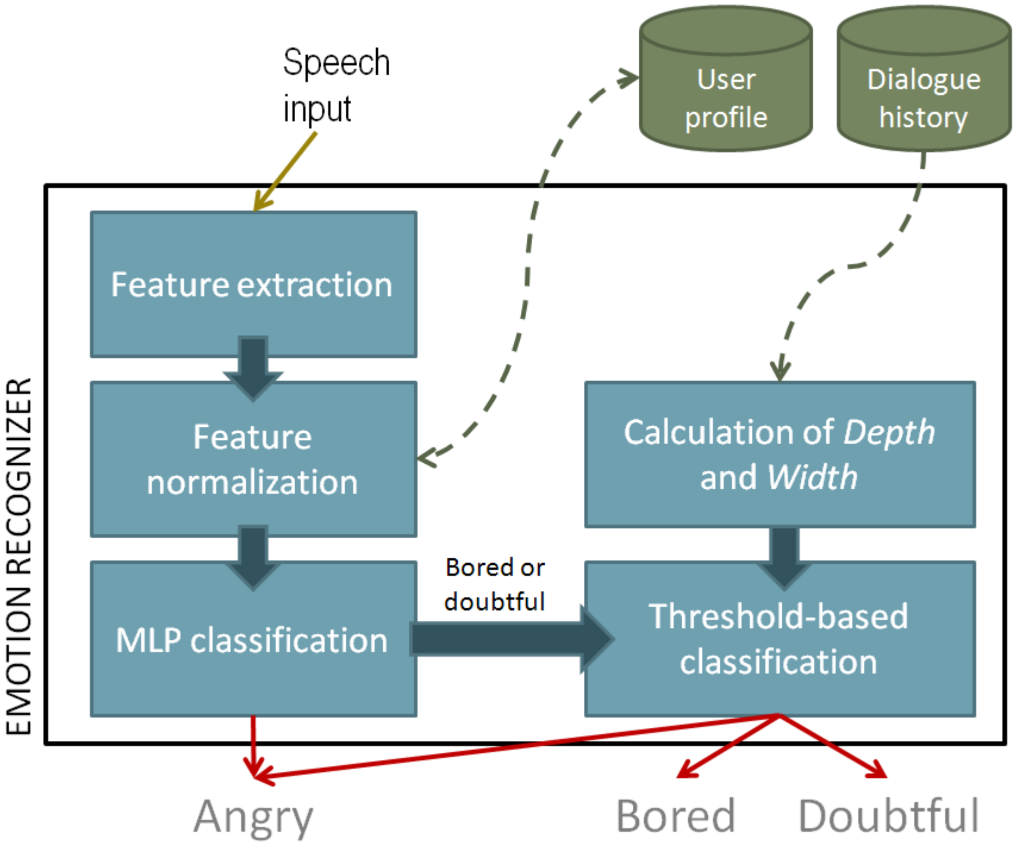


Figure 2: Schema of the emotion recognizer

As can be observed in the figure, the emotion recognizer always chooses one of the three negative emotions under study, not taking *neutral* into account. This is due to the difficulty of distinguishing neutral from emotional speech in spontaneous utterances when the application domain is not highly affective. This is the case of most information providing spoken dialogue systems, for example the UAH system, which we have used to evaluate our proposal (Section 4), in which 85% of the utterances are neutral. Thus, a baseline algorithm which always chooses “neutral” would have a very high accuracy (in our case 85%), which is difficult to improve by classifying the rest of emotions, that are very subtlety produced.

Instead of considering neutral as another emotional class, we calculate the most likely non-neutral category and then the dialogue manager employs the intention information together with this category to decide whether to take the user input as emotional or neutral, as will be explained in Section 5.

The first step for emotion recognition is **feature extraction**. The aim is to compute features from the speech input which can be relevant for the detection of emotion in the user’s voice. We extracted the most representative selection from the list of 60 features shown in Table 1. The feature selection process is carried out from a corpus of dialogues on demand, so that when new dialogues are available, the selection algorithms can be executed again and the list of representative features can be updated. The features are selected by majority voting of a forward selection algorithm, a genetic search, and a ranking filter using the default values of their respective parameters provided by Weka (Witten and Frank, 2005).

Groups	Features	Physiological changes related to emotion
Pitch	Minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation coefficient, slope, and error of the linear regression.	Tension of the vocal folds and the sub glottal air pressure.
First two formant frequencies and their bandwidths	Minimum value, maximum value, range, mean, median, standard deviation and value in the first and last voiced segments.	Vocal tract resonances.
Energy	Minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation, slope, and error of the energy linear regression.	Vocal effort, arousal of emotions.
Rhythm	Speech rate, duration of voiced segments, duration of unvoiced segments, duration of longest voiced segment and number of unvoiced segments.	Duration and stress conditions.
References	*(Hansen, 1996), (Ververidis and Kotropoulos, 2006) (Morrison et al., 2007), (Batliner et al., 2011)*	

Table 1: Features employed for emotion detection from the acoustic signal

The second step of the emotion recognition process is **feature normalization,** with which the features extracted in the previous phase are normalized around the user neutral speaking style. This enables us to make more representative classifications, as it might happen that a user ‘A’ always speaks very fast and loudly, while a user ‘B’ always speaks in a very relaxed way. Then, some acoustic features may be the same for ‘A’ neutral as for ‘B’ angry, which would make the automatic classification fail for one of the users if the features are not normalized.

The values for all features in the neutral style are stored in a user profile. They are calculated as the most frequent values of the user previous utterances which have been annotated as neutral. This can be done when the user logs in to the system before starting the dialogue. If the system does not have information about the identity of the user, we take the first user utterance as neutral assuming that he is not placing the telephone call already in a negative emotional state. In our case, the corpus of spontaneous dialogues employed to train the system (the UAH corpus, to be described in Section 4), does not have login information and thus the first utterances were taken as neutral. For the new user calls of the experiments (Section 5), recruited users were provided with a numeric password.

Once we have obtained the normalized features, we classify the corresponding utterance with a **multilayer perceptron (MLP)** into two categories: *angry* and *doubtful_or_bored.* if an utterance is classified as *angry*, the emotional category is passed to the mental state composer, which merges it with the intention information to represent the current mental state of the user. If the utterance is classified as *doubtful_or_bored*, it is passed through an additional step in which it is classified according to two dialogue parameters: *depth* and *width*. The precision values obtained with the MLP are discussed in detail in (Callejas and López-Cózar, 2008a) where we evaluated the accuracy of the initial version of this emotion recognizer.

Dialogue context is considered for emotion recognition by **calculating *depth* and *width.*** *Depth* represents the total number of dialogue turns up to a particular point of the dialogue, whereas *width* represents the total number of extra turns needed throughout a subdialogue to confirm or repeat information. This way, the recognizer has information about the situations in the dialogue that may lead to certain negative emotions, e.g. a very long dialogue might increase the probability of boredom, whereas a dialogue in which most turns were employed to confirm data can make the user angry.

The computation of *depth* and *width* is carried out according to the dialogue history, which is stored in log files. *Depth* is initialized to 1 and incremented with each new user turn, as well as each time the interaction goes backwards (e.g. to the main menu). *Width* is initialized to 0 and is increased by 1 for each user turn generated to confirm, repeat data or ask the system for help.

Once these parameters have been calculated, the emotion recognizer carries out a **classification based on thresholds** as schematized in Figure 3. An utterance is recognized as *bored* when more than 50% of the dialogue has been employed to repeat or confirm information to the system. The user can also be *bored* when the number of errors is low (below 20%) but the dialogue has been long. If the dialogue has been short and with few errors, the user is considered to be *doubtful* because in the first stages of the dialogue is more likely that users are unsure about how to interact with the system.

Finally, an utterance is recognized as *angry* when the user was considered to be *angry* in at least one of his two previous turns in the dialogue (as with human annotation), or the utterance is not in any of the previous situations (i.e. the percentage of the full dialogue depth comprised by the confirmations and/or repetitions is between 20% and 50%).

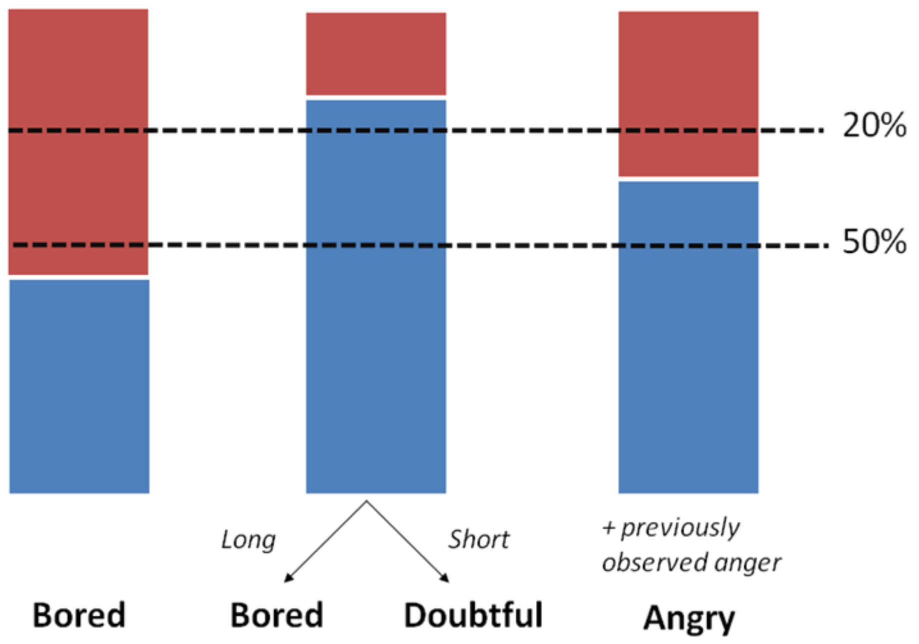


Figure 3: Emotion classification based on dialogue features (blue = depth, red = width)

The thresholds employed are based on an analysis of the UAH emotional corpus, which will be described in Section 4. The computation of such thresholds depends on the nature of the task for the dialogue system under study and how "emotional" the interactions can be.

3.2. The intention recognizer

The methodology that we have developed for modelling the user intention extends our previous work in statistical models for dialogue management (Griol et al., 2008). We define *user intention* as the predicted next user action to fulfil their objective in the dialogue. It is computed taking into account the information provided by the user throughout the dialogue history, and the last system turn.

The formal description of the proposed model is as follows. Let A_i be the output of the dialogue system (the system answer) at time *i*, expressed in terms of dialogue acts. Let U_i be the semantic representation of the user intention. We represent a dialogue as a sequence of pairs (*system-turn, user-turn*)

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

where A_1 is the greeting turn of the system (the first dialogue turn), and U_n is the last user turn.

We refer to the pair $(A_i;U_i)$ as S_i, which is the state of the dialogue sequence at time *i*. Given the representation of a dialogue as this sequence of pairs, the objective of the user intention recognizer at time *i* is to select an appropriate user answer U_i. This selection is a local process for each time *i*, which takes into account the sequence of dialogue states that precede time *i* and the system answer at time *i*. If the most likely user intention level U_i is selected at each time *i*, the selection is made using the following maximization rule:

$$\widehat{U_\iota} = \underbrace{argmax}_{U_i \in U} P(U_i|S_1, \cdots, S_{i-1}, A_i)$$

where the set *U* contains all the possible user answers.

As the number of possible sequences of states is very large, we establish a partition in this space (i.e., in the history of the dialogue up to time *i*). Let UR_i be what we call *user register* at time *i*. The *user register* can be defined as a data structure that contains information about concepts and attributes values provided by the user throughout the previous dialogue history. The information contained in UR_i is a summary of the information provided by the user up to time *i.* That is, the semantic interpretation of the user utterances during the dialog and the information that is contained in the user profile.

The user profile is comprised of user's:

- Id, which he can use to log in to the system.
- Gender.
- Experience, which can be either 0 for novel users (first time the user calls the system) or the number of times the user has interacted with the system.
- Skill level, estimated taking into account the level of expertise, the duration of their previous dialogues and the time that was

necessary to access a specific content and the date of the last interaction with the system. A low, medium, high or expert level is assigned using these measures.

- Most frequent objective of the user.
- Reference to the location of all the information regarding the previous interactions and the corresponding objective and subjective parameters for that user.
- Parameters of the user neutral voice as explained in Section 3.1.

The partition that we establish in this space is based on the assumption that two different sequences of states are equivalent if they lead to the same *UR*. After applying the above considerations and establishing the equivalence relations in the histories of dialogues, the selection of the best U_i is given by:

$$\widehat{U_\imath} = \underbrace{argmax}_{U_i \in U} P(U_i | UR_{i-1}, A_i)$$

To recognize the user intention, we assume that the exact values for the attributes provided by the user are not significant. They are important for accessing the databases and constructing the system prompts. However, the only information necessary to determine the user intention and their objective in the dialog is the presence or absence of concepts and attributes. Therefore, the values of the attributes in the *UR* are coded in terms of three values {0, 1, 2}, where each value has the following meaning:

- *0*: The concept is not activated, or the value of the attribute has not yet been provided by the user.
- *1*: The concept or attribute is activated with a confidence score that is higher than a given threshold (between 0 and 1). The confidence score is provided during the recognition and understanding processes and can be increased by means of confirmation turns.
- *2*: The concept or attribute is activated with a confidence score that is lower than the given threshold.

We propose the use of a classification process to predict the user intention following the previous equation. The classification function can be defined in several ways. We previously evaluated four alternatives: a multinomial naive Bayes classifier, a n-gram based classifier, a classifier based on grammatical inference techniques, and a classifier based on neural networks (Griol et al., 2006, Griol et al., 2008). The accuracy results obtained with these classifiers were respectively 88.5%, 51.2%, 75.7%, and 97.5%. As the best results were obtained using a multilayer perceptron (MLP), we used MLPs as classifiers for these experiments, where the input layer received the current situation of the dialogue, which is represented by the term ($UR_{i\text{-}1}$,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 dialogue.

4. The UAH dialogue system

Universidad Al Habla (UAH - University on the Line) is a spoken dialogue system that provides spoken access to academic information about the Dept. of Languages and Computer Systems at the University of Granada, Spain (Callejas and López-Cózar, 2005; Callejas and López-Cózar, 2008b). The information that the system provides can be classified in four main groups: subjects, professors, doctoral studies and registration, as shown in Table 2. As can be observed, the system asks the user for different pieces of information before producing a response.

A corpus of 100 dialogues was acquired with this system from student telephone calls. The callers were not recruited and the interaction with the system corresponded to the need of the users to obtain academic information. This resulted in a spontaneous Spanish speech dialogue corpus with 60 different speakers. The total number of user turns was 422 and the recorded material has duration of 150 minutes. In order to endow the system with the capability to adapt to the user mental state, we carried out two different annotations of the corpus: intention and emotional annotation.

Category	Information provided by the user (including examples)		Information provided by the system
Subject	*Name*	Compilers	Degree, lecturers, responsible lecturer, semester, credits, web page
	Degree, in case that there are several subjects with the same name	Computer Science	
	Group name and optionally *type*, in case he asks for information about a specific group	A Theory A	Timetable, lecturer
Lecturers	Any combination of *name* and *surnames*	Zoraida Zoraida Callejas Ms. Callejas	Office location, contact information (phone, fax, email), groups and subjects, doctoral courses
	Optionally *semester*, in case he asks for the tutoring hours	First semester Second semester	Tutoring timetable
Doctoral studies	Name of a doctoral program	Software development	Department, responsible
	Name of a course	Object-oriented programming	Type, credits
Registration	Name of the deadline	Provisional registration confirmation	Initial time, final time, description

Table 2: Information provided by the UAH system

Firstly, we estimated the user intention at each user utterance by using concepts and attribute-value pairs. One or more concepts represent the intention of the utterance, and a sequence of attribute-value pairs contains the information about the values provided by the user. We defined four concepts to represent the different queries that the user can perform (*Subject*, *Lecturers*, *Doctoral studies*, and *Registration*), three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*), and eight attributes (*Subject-Name*, *Degree*, *Group-Name*, *Subject-Type*, *Lecturer-Name*, *Program-Name*, *Semester*, and *Deadline*). An example of the semantic interpretation of an input sentence is shown in Figure 4.

User Turn:
I want to get information about Language Processors of Computer Science.
Semantic Representation:
(*Subject*)
    Subject-Name: Language Processors
    Degree: Computer Science

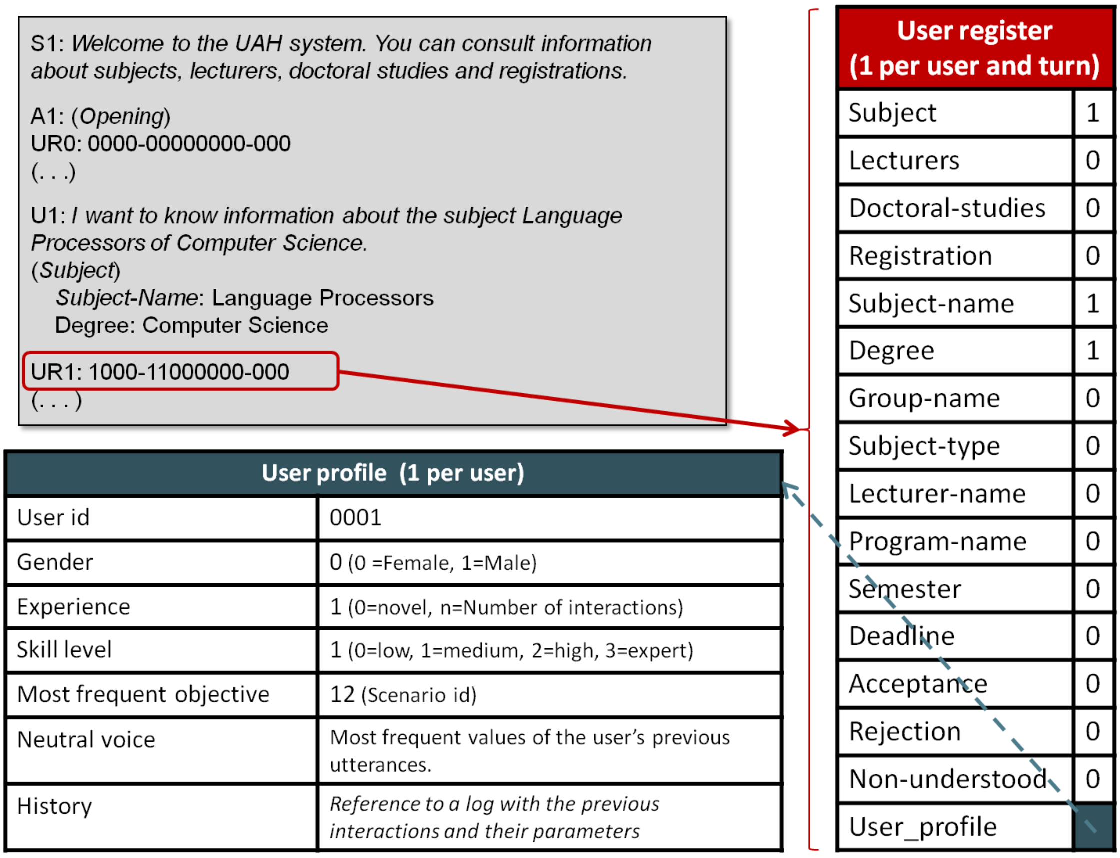
Figure 4: Example of the semantic interpretation of a user utterance with the UAH system

The labelling of the system turns is similar to the labelling defined for the user turns. To do so, 30 task-dependent concepts were defined:

- Task-independent concepts (*Affirmation*, *Negation*, *Not-Understood*, *New-Query*, *Opening*, and *Closing*).
- Concepts used to inform the user about the result of a specific query (*Subject*, *Lecturers*, *Doctoral-Studies*, and *Registration*).
- Concepts defined to require the user the attributes that are necessary for a specific query (*Subject-Name*, *Degree*, *Group-Name*, *Subject-Type*, *Lecturer-Name*, *Program-Name*, *Semester*, and *Deadline*).
- Concepts used for the confirmation of concepts (*Confirmation-Subject*, *Confirmation-Lecturers*, *Confirmation-DoctoralStudies*, *Confirmation-Registration*) and attributes (*Confirmation-SubjectName*, *Confirmation-Degree*, *Confirmation-GroupName*, *Confirmation-SubjectType*, *Confirmation-LecturerName*, *Confirmation-ProgramName*, *Confirmation-Semester*, and *Confirmation-Deadline*).

The *UR* defined for the task is a sequence of 16 fields, corresponding to the four concepts (*Subject*, *Lecturers*, *Doctoral-Studies*, and *Registration*), eight attributes (*Subject-Name*, *Degree*, *Group-Name*, *Subject-Type*, *Lecturer-Name*, *Program-Name*, *Semester*, and *Deadline*) defined for the task, the three task-independent concepts that the users can provide (*Acceptance*, *Negation*, and *Not-Understood*), and a reference to the user profile.

Using the codification previously described for the information in the UR, every dialog begins with a dialog register in which every value is equal to 0 and the greeting turn of the system. Each time the user provides information, it is used to update the previous UR and obtain the current one, as shown in Figure 5. If there is information available about the user gender, usage statistics and skill level, it is incorporated to a user profile that is addressed from the user register, as was explained in Section 3.2.



S1: *Welcome to the UAH system. You can consult information about subjects, lecturers, doctoral studies and registrations.*

A1: (*Opening*)
UR0: 0000-00000000-000
(. . .)

U1: *I want to know information about the subject Language Processors of Computer Science.*
(*Subject*)
Subject-Name: Language Processors
Degree: Computer Science

UR1: 1000-11000000-000
(. . .)

User profile (1 per user)	
User id	0001
Gender	0 (0 =Female, 1=Male)
Experience	1 (0=novel, n=Number of interactions)
Skill level	1 (0=low, 1=medium, 2=high, 3=expert)
Most frequent objective	12 (Scenario id)
Neutral voice	Most frequent values of the user's previous utterances.
History	*Reference to a log with the previous interactions and their parameters*

User register (1 per user and turn)	
Subject	1
Lecturers	0
Doctoral-studies	0
Registration	0
Subject-name	1
Degree	1
Group-name	0
Subject-type	0
Lecturer-name	0
Program-name	0
Semester	0
Deadline	0
Acceptance	0
Rejection	0
Non-understood	0
User_profile	

Figure 5: Excerpt of a dialogue with its correspondent user profile and user register for one of the turns

Secondly, we assigned an emotion category to each user utterance. Our main interest was to study negative user emotional states, mainly to detect frustration because of system malfunctions. To do so, the negative emotions tagged were *angry*, *bored* and *doubtful* (in addition to neutral). Nine annotators tagged the corpus twice and the final emotion assigned to each utterance was the one annotated by the majority of annotators. A detailed description of the annotation of the corpus and the intricacies of the calculation of inter-annotator reliability can be found in (Callejas and López-Cózar, 2009).

5. Evaluation methodology

To evaluate the proposed model for predicting the user mental state discussed in Section 3, we have developed an enhanced version of the UAH system in which we have included the module shown in Figure 1.

Additionally, we have modified the dialogue manager to process mental state information in order to reduce the impact of the user negative states on the communication and the user experience, by adapting the system responses considering mental states. The dialogue manager tailors the next system answer to the user state by changing the help providing mechanisms, the confirmation strategy and the interaction flexibility. The conciliation strategies adopted are, following the constraints defined in (Burkhardt et al., 2009), straightforward and well delimited in order not to make the user loose the focus on the task. They are as follows:

- If the recognized emotion is *doubtful* and the user has changed his behaviour several times during the dialogue, the dialogue manager changes to a system-directed initiative and adds at the end of each prompt a help message describing the available options. This approach is also selected when the user profile indicates that the user is non-expert (or if there is no profile for the current user), and when his first utterances are classified as doubtful.

- In the case of *anger*, if the dialogue history shows that there have been many errors during the interaction, the system apologizes and switches to DTMF (Dual-Tone Multi-Frequency) mode. If the user is assumed to be angry but the system is not aware of any error, the system's prompt is rephrased with more agreeable phrases and the user is advised that they can ask for help at any time.

- In the case of *boredom*, if there is information available from other interactions of the same user, the system tries to infer from those dialogues what the most likely objective of the user might be. If the detected objective matches the predicted intention, the system takes the information for granted and uses implicit confirmations. For example, if a student always asks for subjects of a certain degree, the system can directly disambiguate a subject if it is in several degrees.

- In any other case, the emotion is assumed to be *neutral*, and the next system prompt is decided only on the basis of the user intention and the user profile (i.e., considering his preferences, previous interactions, and expertise level).

In order to evaluate the benefits of including the mental state prediction in the system, we have employed a user simulator to gather a corpus of new dialogues that allows obtaining a more detailed study with a higher range of emotional behaviours. Additionally, we have recorded a corpus of 150 dialogues with 6 recruited users to evaluate the system in more realistic conditions and to gather subjective judgments about it. Figure 6 presents a schematic representation of the corpora used and the users that recorded the dialogues.

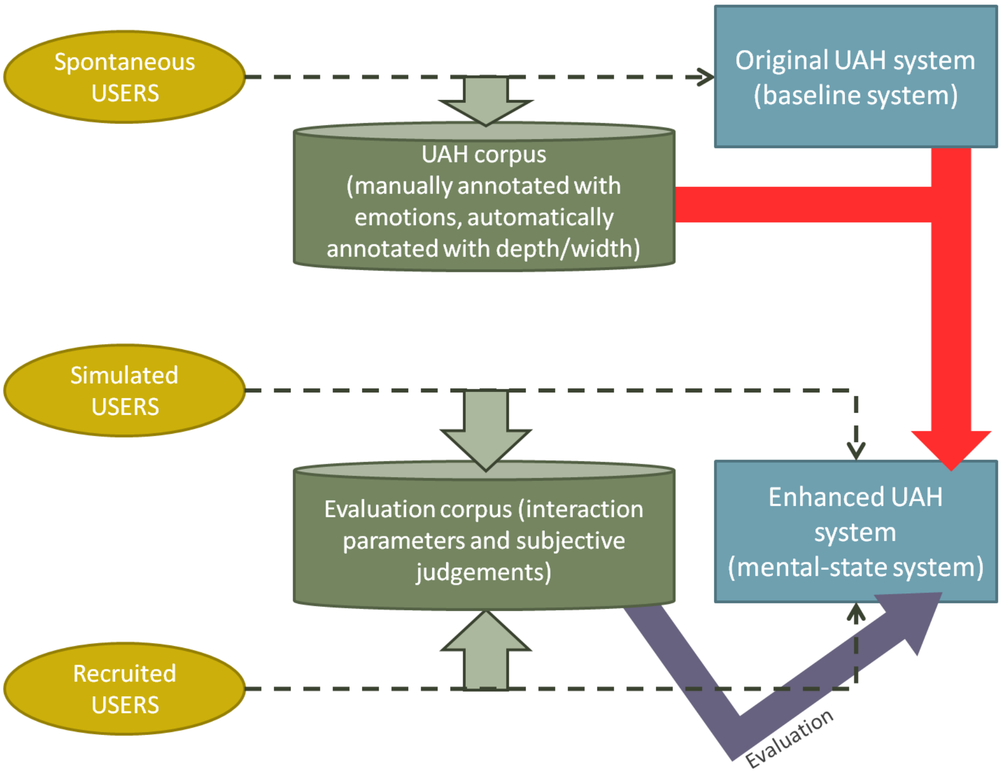

Spontaneous USERS
Original UAH system (baseline system)
UAH corpus (manually annotated with emotions, automatically annotated with depth/width)
Simulated USERS
Evaluation corpus (interaction parameters and subjective judgements)
Enhanced UAH system (mental-state system)
Recruited USERS
Evaluation

Figure 6: Scheme of the corpora used in the paper

5.1. Evaluation with a user simulator

User simulators make it possible to generate a large number of dialogues in a very simple way, reducing the time and effort that would be needed for the detailed evaluation of the quality of the services provided by a dialogue system (López-Cózar et al., 2006). With this aim, we had previously developed a technique which we have successfully applied to the simulation of other systems in the domains of help-desk assistance, railway information, booking facilities and health-care (Griol et al., 2009b; Griol et al., 2010). This simulator carries out the functions of the ASR and NLU modules. An additional error simulator module is used to perform error generation and the addition of ASR confidence scores (Griol et al., 2007). The number of errors that are introduced in the recognized sentence can be modified to adapt the error simulator module to the operation of any ASR and NLU modules.

For these experiments, we have adapted this simulator to generate simulated user intentions following the semantics of the UAH system. As in the intention recognizer, the user simulation generates the user intention level, that is, the user simulator provides concepts and attributes that represent the intention of the user utterance. Additionally, we have added as a novel function the simulation of the output of the emotion recognizer. In order to do so, the selection of the possible users' emotions coincides with the set described for the development of our emotion recognizer for the system (boredom, anger, doubtfulness and neutral).

To generate the emotion label for each turn of the simulated user, we employ the rule-based approach shown in Figure 7, which is based on dialogue information similar to the threshold method employed as a second step in the emotion recognizer described in Section 3. In each case, the method chooses randomly (0.5 probabilities) between an emotion (doubtful, bored or angry) and neutral. The probability of choosing the emotion rises to 0.7 when the same emotion was chosen in the previous turn, which allows simulating moderate changes of the emotional state. Although the simulated users resemble the behaviour of the real users of the UAH corpus (the changes in the emotional state correspond to the same transitions observed in the dialogue states), they are more emotional, as the probability of neutral in the corpus was 0.85. This way, it is possible to obtain different degrees of emotional behaviour with which to evaluate the benefits of our proposal.

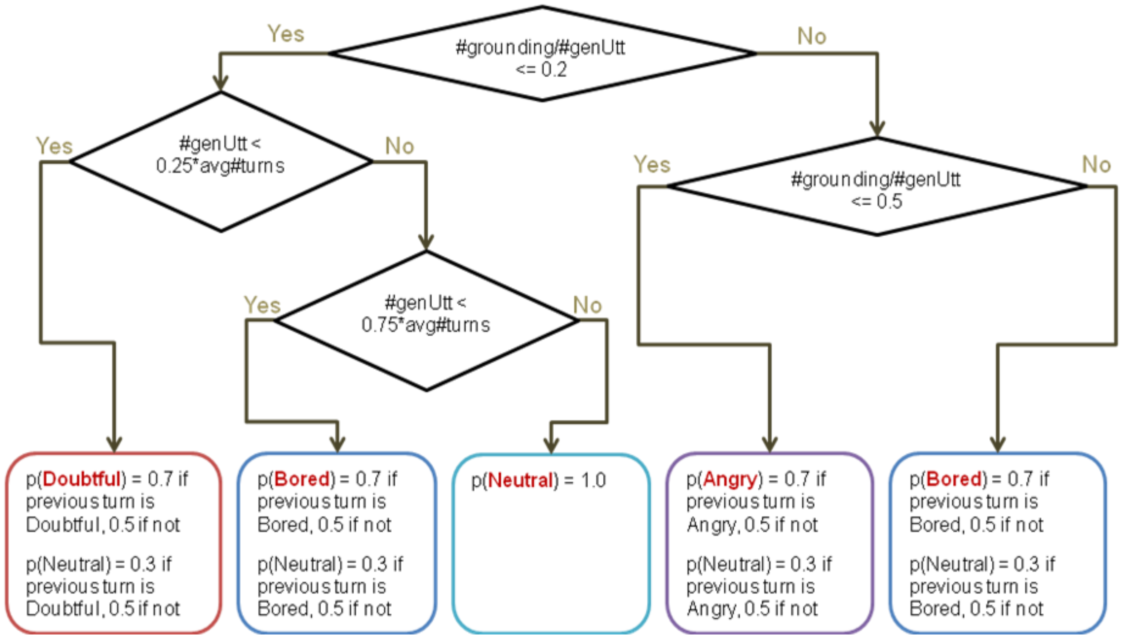

#grounding/#genUtt <= 0.2
Yes
No
#genUtt < 0.25*avg#turns
Yes
No
#grounding/#genUtt <= 0.5
Yes
No
#genUtt < 0.75*avg#turns
Yes
No
p(Doubtful) = 0.7 if previous turn is Doubtful, 0.5 if not
p(Neutral) = 0.3 if previous turn is Doubtful, 0.5 if not
p(Bored) = 0.7 if previous turn is Bored, 0.5 if not
p(Neutral) = 0.3 if previous turn is Bored, 0.5 if not
p(Neutral) = 1.0
p(Angry) = 0.7 if previous turn is Angry, 0.5 if not
p(Neutral) = 0.3 if previous turn is Angry, 0.5 if not
p(Bored) = 0.7 if previous turn is Bored, 0.5 if not
p(Neutral) = 0.3 if previous turn is Bored, 0.5 if not

Figure 7: Process for emotion generation for each turn of the user simulator (#genUtt = number of utterances generated so far in the dialogue, #grounding = number of utterances corresponding to grounding actions, avg#turns = average number of turns of the generated dialogues)

A user request for closing the dialogue is selected once the system has provided the information defined in the objective(s) of the dialogue. The dialogues that fulfil this condition before a maximum number of turns are considered successful. The dialogue manager considers that the dialogue is unsuccessful and decides to abort it when the following conditions hold:

- The dialogue exceeds the maximum number of user turns, specified taking into account real dialogues for the task.

- The answer selected by the dialog manager corresponds with a query not required by the user simulator.

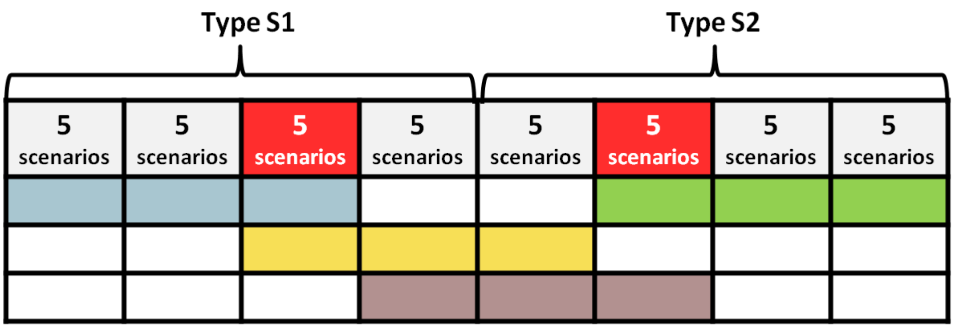
- The database query module generates an error warning because the user simulator has not provided the mandatory information needed to carry out the query.
- The oral response generator generates an error when the selected answer involves the use of a data not provided by the user simulator.

The user simulation technique was used to acquire a total of 2000 successful dialogues, both including and not including the prediction module of the mental state in the architecture of the system (i.e., 1000 dialogues using the architecture shown in Figure 1, and 1000 dialogues without including the described mental state prediction module).

A set of 40 scenarios were manually defined to consider the different queries that may be performed by users. Two main types of scenario were specified. Scenarios of type S1 defined only one objective for the dialogue (e.g., to obtain timetable information of a specific subject). Scenarios of type S2 defined two objectives for the dialogue (e.g. to obtain timetables of a specific subject and registration deadlines for the corresponding degree).

5.2. Evaluation with real users

Additionally, we evaluated the behaviour of the mental-state version of the UAH system with six recruited users using the same set of type S1 and S2 scenarios designed for the user simulation. Four of them recorded 30 dialogues (15 scenarios with the baseline system and 15 with the mental-state system), and two of them recorded 15 dialogues (15 dialogues with the baseline or the mental-state system only). Thus, as shown in Figure 8, a total of 150 dialogues were recorded in such a way that there were two dialogues recorded per scenario, three in the case of the five most frequent scenarios of each type as observed in the UAH corpus.



(15 scenarios/system x 1 dialogue/scenario x 2 systems) x 2 users = 60 dialogues
(15 scenarios/system x 1 dialogue/scenario x 2 systems) x 2 users = 60 dialogues
+ (15 scenarios/system x 1 dialogue/scenario x 1 system) x 2 users = 30 dialogues

150 dialogues

🟦 Scenarios by user1 and user2 with *baseline* and by user3 and user4 with *mental_state.*
🟩 Scenarios by user3 and user4 with *baseline* and by user1 and user2 with *mental_state.*
🟨 Scenarios by user5 with *baseline.*
🟫 Scenarios by user6 with *mental_state.*
🟥 Five most frequent scenarios of Type Si.

Figure 8: Acquisition of dialogues with recruited users for the evaluation of our proposal

5.3. Evaluation metrics

To compare the baseline and mental-state versions of the UAH system (with both the simulated and recruited users) we computed the mean value for the evaluation measures shown in Table 3, which we extracted from different studies (Ai et al., 2007, Griol et al. 2009a, Schatzmann et al. 2005). We then used two-tailed t-tests to compare the means across the different types of scenarios and users as described in (Ai et al., 2007). The significance of the results discussed in Section 6 was computed using the SPSS software with a significance level of 95%[1].

[1] The degrees of freedom that SPSS employs for t-tests are N-1 in case the compared groups have the same number of samples (N), and N1+N2-1 when they differ in the number of samples (N1 and N2). In these experiments, the degrees of freedom were 1,074 when comparing the baseline and *mental state* system (N=1075) and 2,149 when comparisons were carried out between the simulated and the recruited users (N1=2000 and N2=150 respectively).

Dialogue success
Dialogue success rate (%success). The percentage of successfully completed tasks. In each scenario, the user has to obtain one or several pieces of information, and the dialogue success depends on whether the system provides the correct data (according to the aims of the scenario) or incorrect data to the user.
Average number of corrected errors per dialogue (nCE). The average of errors detected and corrected by the dialogue manager. We have considered only the errors that modify the values of the attributes and that could cause dialogue failure.
Average number of uncorrected errors per dialogue (nNCE). The average of errors not corrected by the dialogue manager. Again, only errors that modify the values of the attributes are considered.
Error correction rate (%ECR) is the percentage of corrected errors, computed as nCE/(nCE + nNCE).
High-level dialogue features
Average number of turns per dialogue (avg#turns/dial).
Percentage of different dialogues (%diff).
Number of repetitions of the most seen dialogue (#repMS).
Number of turns of the most seen dialogue (#turnsMS).
Number of turns of the shortest dialogue (#turnsSh).
Number of turns of the longest dialogue (#turnsLo).
Ratio users vs. system actions (us/sysAct).
Dialogue style/cooperativeness measures
System dialogue acts: Confirmation of concepts and attributes, Questions to require information, and Answers generated after a database query.
Confirmation rate (%confirm) was computed as the ratio between the number of explicit confirmations turns (nCT) and the number of turns in the dialogue (nCT/nT).
User dialogue acts: Request to the system, Provide information, Confirmation, Yes/No answers, and Other answers.
Goal directed actions vs. grounding actions: Goal directed actions are requesting and providing information, while grounding actions are explicit and implicit confirmations, dialogue formalities (greetings, instructions, etc.) and unrecognized actions.

Table 3: Evaluation measures based on the interaction parameters gathered from the dialogues of simulated and recruited users.

In addition, we asked the recruited users to complete a questionnaire to assess their subjective opinion about system performance. The questionnaire had five questions:

Q1: How well did the system understand you?

Q2: How well did you understand the system messages?

Q3: Was it easy to obtain the requested information?

Q4: Was the interaction rate adequate?

Q5: If the system made errors, was it easy for you to correct them?

The possible answers for the questions were: *Never*, *Seldom*, *Sometimes*, *Usually*, *Always*. All the answers were assigned a numeric value between one and five (in the same order as they appear in the questionnaire).

6. Evaluation results

Table 4 shows the comparison of the different high-level measures for the *mental-state* and *baseline* systems.

Description of the metrics in Table 3	Simulated users				Recruited users			
	Baseline		Mental-state		Baseline		Mental-state	
	S1	S2	S1	S2	S1	S2	S1	S2
%success	78%	66%	87%	76%	87%	83%	97%	95%
nCE	0.76	0.71	0.89	0.84	0.85	0.80	0.91	0.88
nNCE	0.21	0.24	0.09	0.11	0.18	0.20	0.09	0.08
%ECR	79%	75%	91%	88%	82%	80%	92%	91%
avgturn/dial	8.4	14.8	4.7	9.2	9.2	15.1	5.8	10.4
%diff	76%	88%	67%	84%	77%	93%	76%	91%
#repMS	7	3	9	7	5	2	8	4
#turnsMS	2	9	2	7	2	9	2	7
#turnsSh	2	7	2	7	2	7	2	7
#turnsLo	14	20	12	18	12	17	9	15

Table 4: Results of the high-level dialogue features defined for the comparison of the mental-state and UAH baseline systems.

As can be observed, on the one hand the success rate for the *mental-state* system is higher than the baseline. This difference showed a significance value of 0,025 in the two-tailed t-test. On the other hand, although the error correction rates were also improved in absolute values by using the *mental-state* system, this relationship was not significant in the t-test. Both results are explained by the fact that we have not designed a specific strategy to improve the recognition or understanding processes and decrease the error rate, but rather our proposal for adaptation to the user mental state overcomes these problems during the dialogue once they are produced. The absolute numbers in Table 4 indicate that the increment in the success rate is slightly higher for S2 dialogues compared to S1 dialogues regardless of the system, but this difference between dialogue types was not significant in the test.

Regarding the number of dialogue turns, the *mental state* system produced shorter dialogues (with a 0,000 significance value in the t-test when compared to the number of turns of the baseline system). As shown in Table 4, this general reduction in the number of turns is particularized also to the case of the longest, shortest and most seen dialogues for the *mental-state* system. This might be because users have to explicitly provide and confirm more information using the *baseline* system, whereas the mental state system automatically adapted the dialogue to the user and the dialogue history.

The *baseline* dialogues have a higher standard deviation (3.80) given that the proportion of number of turns per dialogue is more disperse. The dialogues gathered with the *mental-state* system have a smaller deviation (3.20) since the successful dialogues are usually those which require the minimum number of turns to achieve the objective(s) predefined for both kinds of scenario.

Also, in the two types of scenario, the dialogues acquired using the simulation technique were shorter than those acquired with real users. This can be due to the restriction defined for a maximum numbers of turns per dialogue in the user simulation. Also, there were more dialogues in which the recruited users asked for more information than strictly required to optimally fulfil their scenarios.

Table 5 sets out the results regarding the percentage of different dialogues obtained. When we considered the dialogues to be different only when a different sequence of user intentions was observed, the percentage was lower using the *mental-state* system, due an increment in the variability of ways in which the users can provide the different data required to the mental-state system. This is consistent with the fact that the number of repetitions of the most observed dialogues is higher for the baseline system. As can be observed in the table, this flexibility has a bigger impact in the case of the S2 scenarios as the users must convey more information to the system. Also, recruited users seemed to benefit in a greater extent from the flexibility of the mental-state system than simulated users. This can be because of the user profile information that was stored in the system, which also takes into account the expertise of the user, as explained in Section 4.

When emotions were also taken into account, i.e. when even with the same sequence of intentions two dialogues were considered different if the emotions observed were different, we obtained a higher percentage of different dialogues in the case of the simulated users. This is because of the more varied emotional behaviour endowed to the simulated users, which was one of the objectives of the user simulation, as described in Section 5.1. However, this difference was low because our mental state recognizer tends to classify utterances as emotional rather than neutral, as described in Section 3.

Percentage of different dialogues	Simulated users				Recruited users			
	Baseline		Mental-state		Baseline		Mental-state	
	S1	S2	S1	S2	S1	S2	S1	S2
Difference at intention level only	76%	88%	67%	84%	77%	93%	76%	91%
Difference at mental-state level (intention+emotion)			83%	97%			81%	95%

Table 5: Percentage of different dialogues obtained

We have previously described the differences between both systems in terms of number of turns. Figure 9 shows that there is also a slight reduction in the number of actions per turn for the dialogues of the *mental-state* system (with a 0.000 significance value in the t-test). S1 scenarios contain 1.3 actions per user turn instead of the 1.5 actions in the *baseline* dialogues, whereas for the S2 scenarios the scores are 1.4 and 1.9, respectively. This is again because the users have to explicitly provide and confirm more information using the *baseline* system.

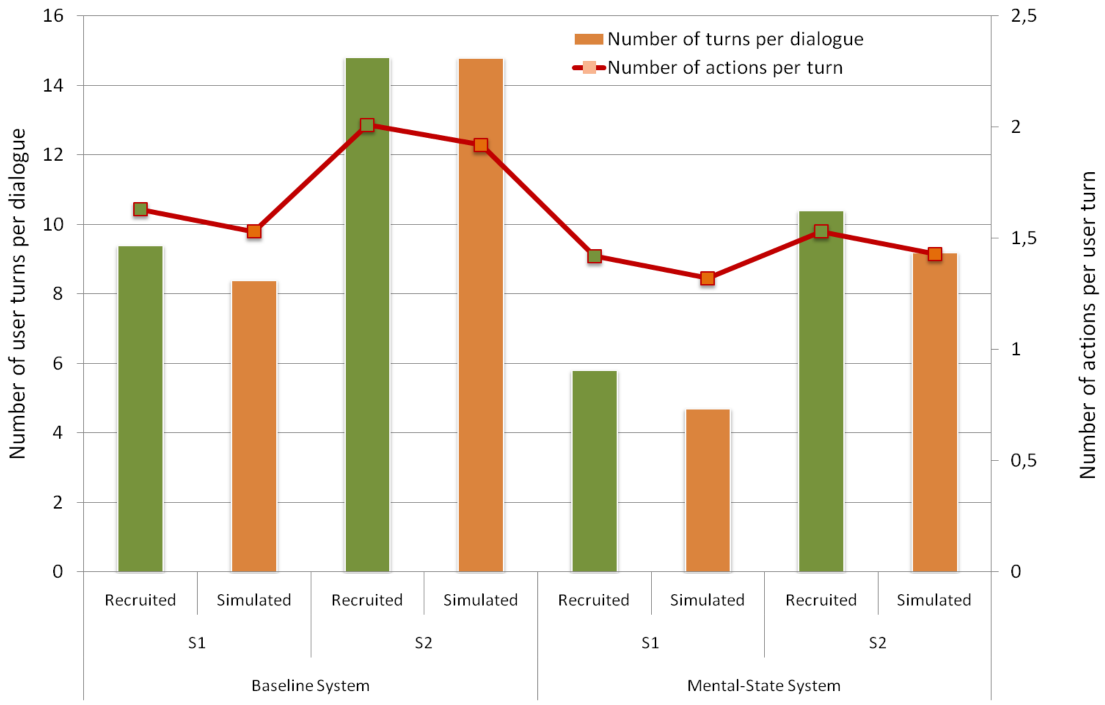


Figure 9: Average number of turns per dialogue and actions per turn in the *Mental-State* and *Baseline* systems

Regarding the dialogue participant activity, Figure 10 shows the ratio of user versus system actions. The dialogues of the *mental-state* system have a higher proportion of system actions due to a reduction of the confirmation turns (0.015 significance). It can be observed only a slight difference in the ration of user/system answers between recruited and simulated users, which was not significant in the t-test.

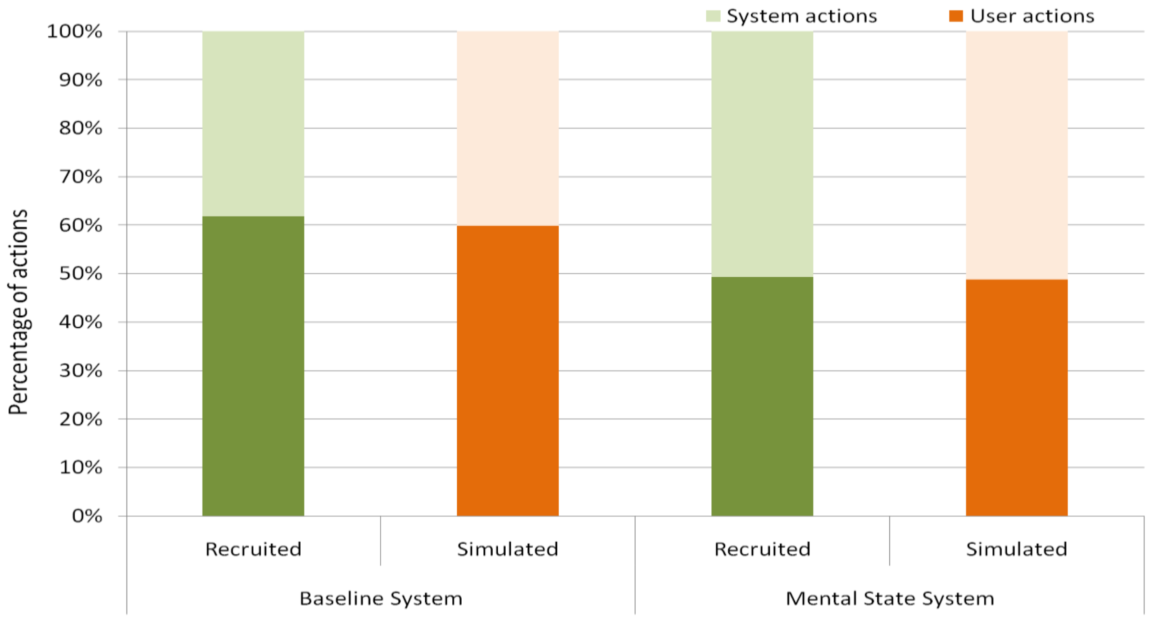


Figure 10: Ratio of user vs. system actions in the *Mental-State* and *Baseline* systems

Regarding dialogue style and cooperativeness, the histograms in Figures 11 and 12 respectively show the frequency of the most dominant user and system dialogue acts in the dialogues collected with the *mental-state* and *baseline* systems. On the one hand, Figure 11 shows that users need to provide less information explicitly using the *mental-state* system; this explains the higher proportion of queries (both differences significant over 98%). It can be observed that there are also only slight differences between the values obtained for both corpora. There is a higher percentage of confirmations and questions in the corpus collected with real users due the higher average number of turns per dialog in this corpus.

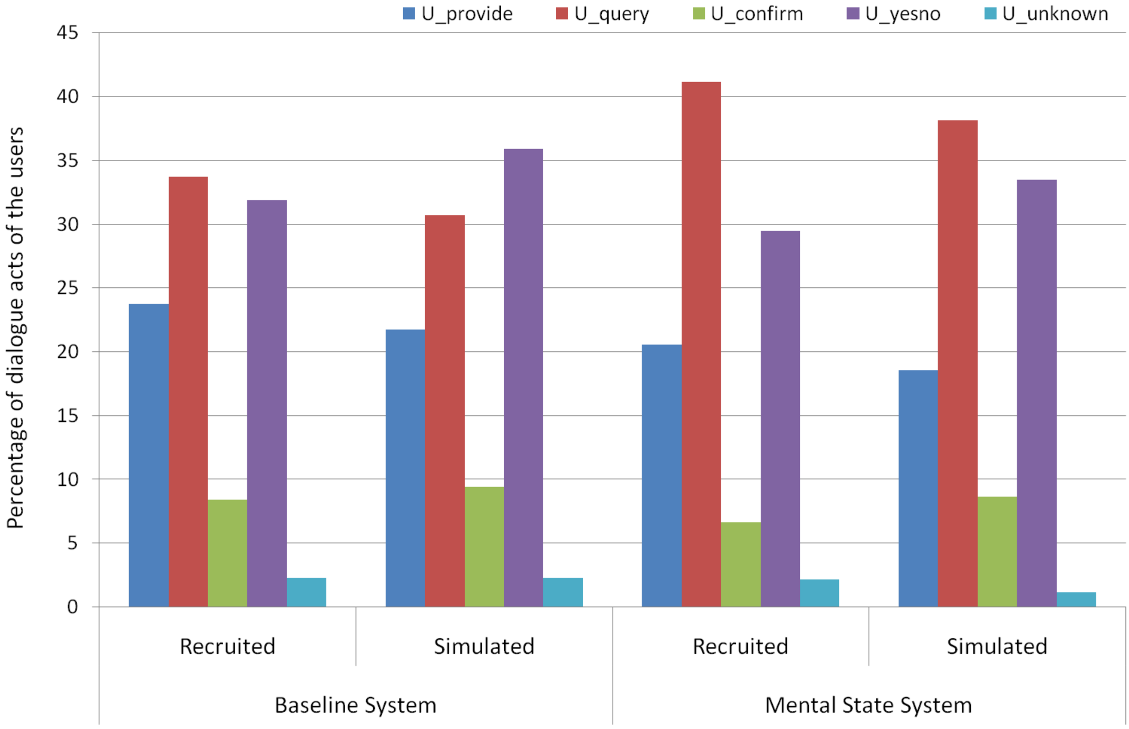


Figure 11: Histogram of user dialogue acts in the *Mental-State* and *Baseline* systems

On the other hand, Figure 12 shows that there is a reduction in the system requests when the *mental-state* system is used. This explains a higher proportion of the inform system action in the *mental-state* system. There was a significant difference between both corpora in the percentage of turns in which the user makes a request to the system. The percentage of this kind of answers is lower in the corpus acquired with real users. This can be explained by the fact that it is less probable that simulated users provide useless information. In fact, there is a lower percentage of users' turns classified as “Other answers”.

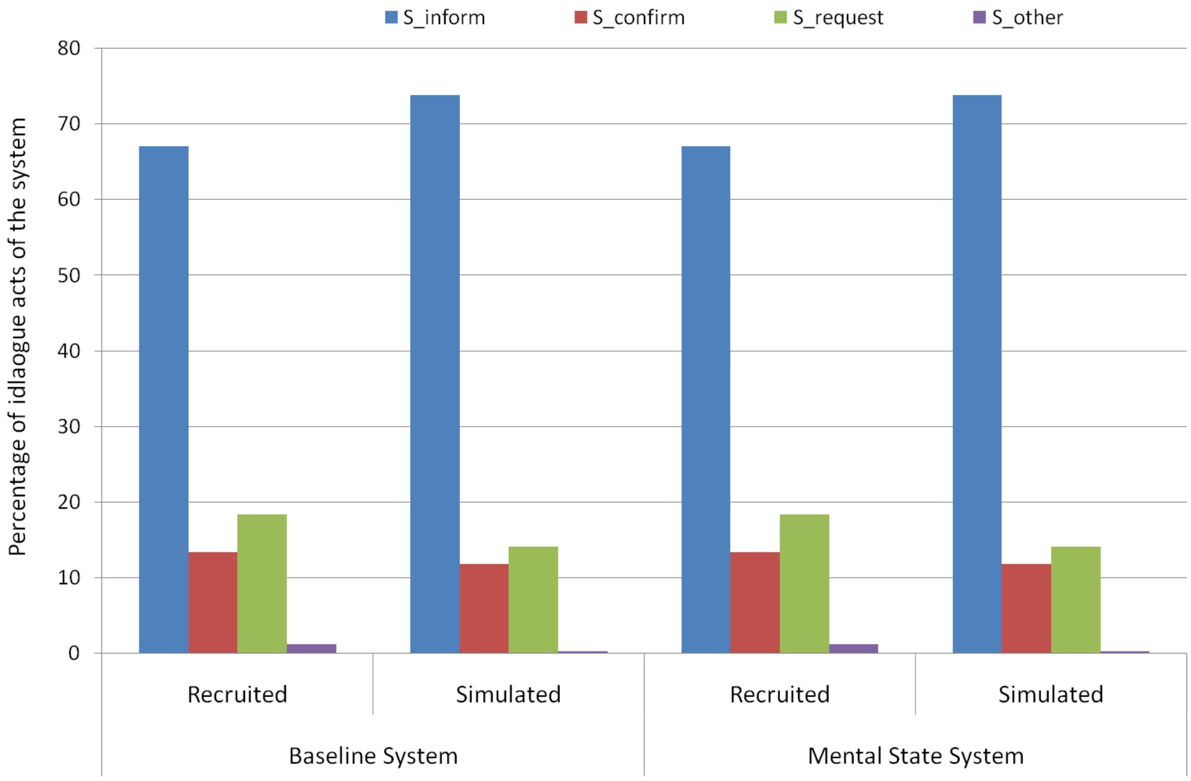


Figure 12: Histogram of system dialogue acts in the *Mental-State* and *Baseline* systems

Additionally, we grouped all user and system actions into three categories: “goal directed” (actions to provide or request information), “grounding” (confirmations and negations), and “rest”. Figure 13 shows a comparison between these categories. As can be observed, the dialogues provided by the *mental-state* system have a better quality, as the proportion of goal-directed actions is higher.

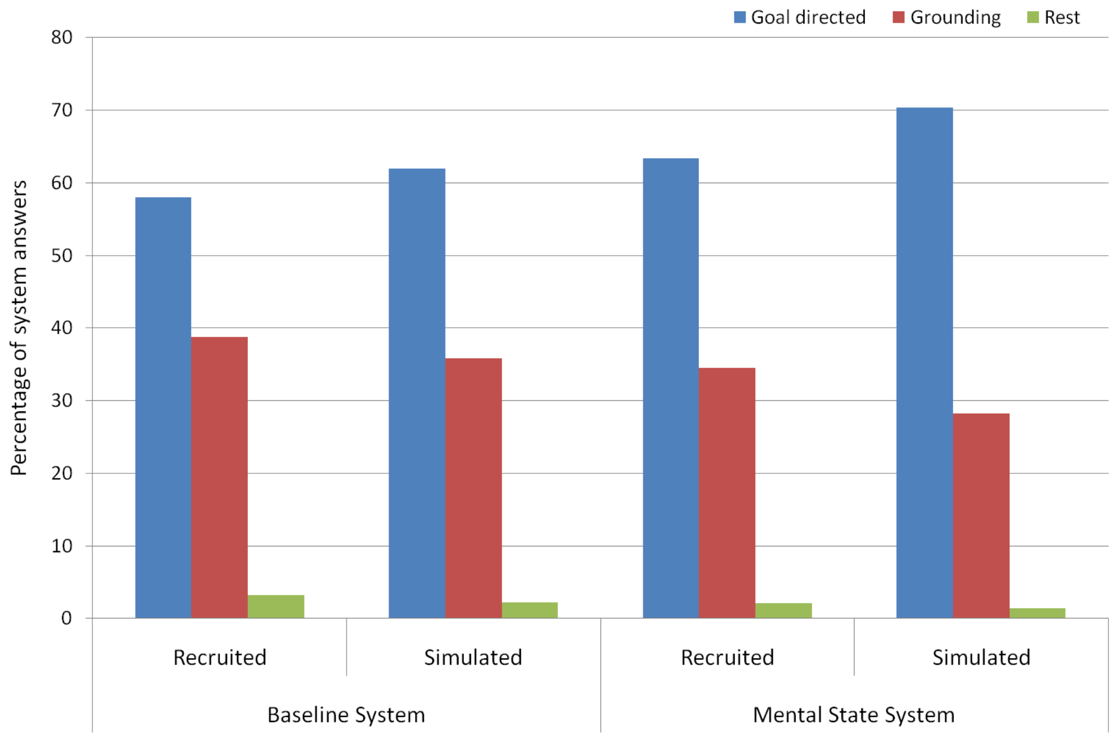


Figure 13: Proportion of turns of goal directed actions, ground actions and rest of possible actions in the *Mental-State* and *Baseline* systems

Table 6 shows the average results obtained with respect to the subjective evaluation carried out by the recruited users. As can be observed, both systems correctly understand the different user queries and obtain a similar evaluation regarding the user observed easiness in correcting errors made by the ASR module. However, the *mental-state* system has a higher evaluation rate regarding the user observed easiness in obtaining the data required to fulfil the complete set of objectives defined in the scenario, as well as the suitability of the interaction rate during the dialogue

	Q1	Q2	Q3	Q4	Q5
Baseline System	4.6	3.6	3.8	3.4	3.2
Mental-State System	4.8	3.9	4.3	4.2	3.3

Table 6: Results of the subjective evaluation of the *Mental-State* and *Baseline* systems with real users (0=worst, 5=best evaluation)

7. Conclusions and future work

In this paper we have presented a method for predicting user mental states in spoken dialogue systems. These states are defined as the combination of the user emotional state and the predicted intention according to their objective in the dialogue. We have proposed an architecture in which our method is implemented as a module comprised of an emotion recognizer and an intention recognizer. The emotion recognizer obtains the user emotional state from the acoustics of his utterance as well as the dialogue history. The intention recognizer decides the next user action and their dialogue goal using a statistical approach that relies on the previous user input and system prompt.

We have evaluated the method with the UAH spoken dialogue system, implementing the mental state prediction module between the natural language understanding module and the dialogue manager. Additionally, we have enhanced the UAH system to deal with the mental state information. In order to do so, we have improved the dialogue manager to take this information into account in order to compute and adapt the system responses.

The evaluation was carried out using a corpus of interactions between the system and an affective user simulator, and also with the interaction of real users with the mental state version of the system. The results show that the improved version of the system performs better in terms of duration of the dialogues, number of turns needed to succeed in the dialogue and number of confirmations and repetitions needed. Additionally, the users judged the system to be better when it could adapt its behaviour to their mental state.

As a future work we plan to annotate the emotions of the corpus collected with real users interacting with the mental state version of the system in order to refine the adaptation strategies of the dialogue manager. Using this corpus we will be able to evaluate the impact of the adapted dialogue management strategies, not only on the performance of the interaction and the subjective experience of the user, but also on the emotional state of the user. This way, we will check whether the adapted strategies can guide the users out of negative emotional states. Also, the annotated corpus, augmented with new dialogues, will offer us the possibility to employ stochastic approaches for optimized dialogue strategies tailored to the user mental states.

Moreover, we are interested in studying how to evaluate and optimize the proposed mental state simulator. For the research presented in the paper, we have used the simulator in order to obtain more emotional dialogues with which to better analyze the benefits of our proposal, a study on the evaluation of the simulator itself constitutes a very challenging possibility for future work.

Acknowledgements

This research has been funded by the Spanish Ministry of Science and Innovation, under the project ASIES TIN2010-17344: Adaptation in Smart Environments and Social Networks to Assist People with Special Needs.

References

Ábalos, N, Espejo, G., López-Cózar, R., Callejas, Z., Griol, D. (2010). A Multimodal Dialogue System for an Ambient Intelligent Application in Home Environments. Lectures Notes in Artificial Intelligence 6231, pp. 484-491.

Acosta, J.C., Ward, N.G. (2009). Responding to user emotional state by adding emotional coloring to utterances. In: Proc. of 10th Annual Conference of the International Speech Communication Association (Interspeech 09). Brighton, United Kingdom, pp. 1587-1590.

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

Baker, R.S.J.d., D'Mello, S.K.D., Rodrigo, M.M.T., Graesser, A.C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies, 68(4), 223-241.

Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E. (2006). A taxonomy of applications that utilize emotional awareness. In: Proc. of the 1st International Language Technologies Conference (IS-LTC 06). Ljubljana, Slovenia, pp. 246-250.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N. (2011). Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech. Computer Speech and Language 25(1), pp. 4-28.

Beun, R.-J. (1994). Mental state recognition and communicative effects. Journal of Pragmatics 21, pp. 191 – 214.

Bickmore, T., Giorgino, T., (2004). Some novel aspects of health communication from a dialogue systems perspective. In: Proc. of AAAI Fall Symposium on Dialogue Systems for Health Communication. Washington DC, USA, pp. 275–291.

Boril, H., Hansen, J.H.L. (2010). Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Transactions on Audio, Speech, and Language Processing 28(6), pp. 1379-1393.

Boril, H., Sadjadi, O., Kleinschmidt, T., Hansen, J. H. L. (2010). Analysis and Detection of Cognitive Load and Frustration in Drivers' Speech. In: Proc. of Interspeech'10. Makuhari, Chiba, Japan, pp. 502-505.

Bosma, W., Andre, E. (2004). Exploiting emotions to disambiguate dialogue acts. In: Proc. of 9th International Conference on Intelligent User Interface. Funchal, Portugal, pp. 85-92.

Bui, T., Poel, M., Nijholt, A., Zwiers, J. (2009) A tractable hybrid DDN-POMDP approach to affective dialogue modeling for probabilistic frame-based dialogue systems. Natural Language Engineering 15(2), pp. 273-307.

Burkhardt, F., van Ballegooy, M., Engelbrecht, K.P., Polzehl, T., Stegmann, J. (2009) Emotion detection in dialog systems – Usecases, strategies and challenges. In: Proc. of International Conference on Affective Computing and Intelligent Interaction (ACII 09). Amsterdam, The Netherlands.

Callejas, Z., López-Cózar, R, (2005) Implementing modular dialogue systems: a case study. In: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE 05). Aalborg, Denmark.

Callejas, Z., López-Cózar, R., (2008a) Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50(5), pp. 416-433.

Callejas, Z., López-Cózar, R., (2008b) Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication 50(8-9), pp. 646-665.

Callejas, Z., López-Cózar, R., (2009) Improving acceptability assessment for the labeling of affective speech corpora. In: Proc. of 10th Annual Conference of the International Speech Communication Association (Interspeech 09). Brighton, United Kingdom, pp. 2863-2866.

Callejas, Z., López-Cózar, R., Ábalos, N., Griol, D. (2011) Affective conversational agents: the role of personality and emotion in spoken interactions. In: D. Pérez-Martín, I. Pascual-Nieto (Eds.) "Conversational Agents and Natural Language Interaction: Techniques and Effective Practices", IGI Global Publishers.

Das, K., Rizzuto, D., Nenadic, Z. (2009). Mental State Estimation for Brain-Computer Interfaces. IEEE Transactions on Biomedical Engineering 56, pp. 2114 -2122.

Delaborde, A., Devillers, L. (2010). Use of non-verbal speech cues in social interaction between human and robot: emotional and interactional markers. In: Proc. Of 3rd International Workshop on Affective Interaction in Natural Environments. Firenze, Italy, pp 75-80.

Dragoni, A. F. (2008). Mental states as multi-context systems. Annals of Mathematics and Artificial Intelligence 54, pp. 265-292.

Dyer, J. R., Shatz, M., Wellman, H. M. (2000). Young children's storybooks as a source of mental state information. Cognitive Development 15, pp. 17 – 37.

Evanini, K., Hunter, P., Liscombe, J., Suendermann, D., Dayanidhi, K., Pieraccini, R. (2008). Caller experience: a method for evaluating dialog systems and its automatic prediction. In: Proc. of the 2008 Spoken Language Technology Workshop (SLT 08). Goa, India, pp. 129-132.

Fairclough, S. H. (2009). Fundamentals of physiological computing. Interacting with Computers 21, pp. 133 – 145.

Ginzburg, J. (1996). Dynamics and the semantics of dialogue. In: Seligman, J., Westerstahl, D. (Eds.), Logic, language and computation vol 1. CSLI Publications.

Gnjatovic, M., Rösner, D. (2008). Adaptive dialogue management in the NIMITEK prototype system. Lecture Notes in Computer Science 5078, pp. 14-25.

Griol, D., Hurtado, L. F., Sanchis, E., Segarra, E. (2006). Managing Unseen Situations in a Stochastic Dialog Model. In: Proc. of AAAI Workshop on Statistical and Empirical Approachesfor Spoken Dialogue Systems. Antwerp, Belgium, pp. 25-30.

Griol, D., Hurtado, L. F., Sanchis, E., Segarra, E. (2007). Acquiring and evaluating a dialog corpus through a dialog simulation technique. In: Proc. of the 8th Annual SIGdial Meeting on Discourse and Dialogue. Antwerp, Belgium, pp. 29-42.

Griol, D., Hurtado, L. F., Segarra, E., Sanchis, E. (2008) A statistical approach to spoken dialog systems design and evaluation. Speech Communication 50(8-9), pp. 666-682.

Griol, D., Callejas, Z., López-Cózar, R. (2009a). A comparison between dialog corpora acquired with real and simulated users. In: Proc. of the 10th Annual SIGdial Meeting on Discourse and Dialogue. London, United Kingdom, pp. 326-332.

Griol, D., Riccardi, G. Sanchis, E. (2009b) A Statistical Dialog Manager for the LUNA Project. In: Proc. of 10th Annual Conference of the International Speech Communication Association (Interspeech 09). Brighton, United Kingdom, pp. 272-275.

Griol, D., McTear M. F., Callejas, Z., López-Cózar, R., Ábalos, N., Espejo, G. (2010). A methodology for learning optimal dialog strategies. Lectures Notes in Artificial Intelligence 6231, pp. 500-507.

Hansen, J. H. L. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Communication 20 (2), pp. 151–170.

Jokinen, K., 2003. Natural interaction in spoken dialogue systems. In: Proc. of the Workshop Ontologies and Multilinguality in User Interfaces. Crete, Greece, pp. 730–734.

Jokinen, K., Mc Tear, M.F. (2010). Spoken Dialogue Systems. Morgan and Claypool Publishers.

Jonker, C. M., Treur, J. (2002). A dynamic perspective on an agent's mental states and interaction with its environment. In: Proc. of the ACM first international joint conference on Autonomous agents and multiagent systems. Bologna, Italy, pp. 865-872.

Katoh, T., Hara, H., Kinoshita, T., Sugawara, K., Shiratori, N. (1998). Behavior of Agents Based on Mental States. In: Proc. of the 13th International Conference on Information Networking. Tokyo, Japan, pp. 199-204.

Khalifa, O. O., Ahmad, Z. H., Gunawan, T. D. (2007). SMaTTS: Standard Malay Text to Speech System. International Journal of Computer Science 2(4), pp. 285-293.

Lee, L., Harkness, K. L., Sabbagh, M. A., Jacobson, J. A. (2005). Mental state decoding abilities in clinical depression. Journal of Affective Disorders 86, pp. 247 – 258.

Litman, D.J., 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 Communication 48(5), pp. 559-590.

López-Cózar, R., Callejas, Z., McTear, M. F. (2006). Testing the performance of spoken dialogue systems by means of an artificially simulated user. Artificial Intelligence Review 26(4), pp.291-323.

López-Cózar, R., Callejas, Z., Kroul, M., Nouza, J., Silovský, J. (2008). Two-level fusion to improve emotion classification in spoken dialogue systems. Lecture Notes in Computer Science 5246, pp. 617-624.

Lourens, T., van Berkel, R., Barakova, E. (2010). Communicating emotions and mental states to robots in a real time parallel framework using Laban movement analysis. Robotics and Autonomous Systems 58, pp. 1256 – 1265.

Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (Eds.). (2010). Computers Helping People with Special Needs, Proc. 12th International Conference on Computers Helping People with Special Needs (ICCHP 2010), Lecture Notes on Computer Science 4061.

Morrison, D., Wang, R., Silva, L. C. D., 2007. Ensemble methods for spoken emotion recognition in call-centers. Speech Communication 49 (2), pp. 98–112.

Nisimura, R., Omae, S., Kawahara, H., Irino, T. (2006). Analyzing dialogue data for real-world emotional speech classification. Proc. of 9th International Conference on Spoken Language Processing (Interspeech 2006 — ICSLP). Pittsburgh, USA, pp. 1822-1825.

Ohkawa, Y., Suzuki, M., Ogasawara, H., Ito, A., Makino, S. (2009). A speaker adaptation method for non-native speech using learners’ native utterances for computer-assisted language learning systems. Speech communication 51(10), pp. 875-882.

Osatuke, K., Stiles, W. B. (2010). Relationship between mental states in depression: The assimilation model perspective. Psychiatry Research. In Press, Corrected Proof.

Oztop, E., Wolpert, D., Kawato, M. (2005). Mental state inference using visual control parameters. Cognitive Brain Research 22, pp. 129 – 151.

Piccinini, G. (2004). Functionalism, computationalism, and mental states. Studies In History and Philosophy of Science 35, pp. 811 – 833.

Pittermann, J., Pittermann, A., Minker, W. (2010). Emotion recognition and adaptation in spoken dialogue systems. International Journal of Speech Technology 13, pp. 49-60.

Riccardi, G., Hakkani-Tür, D. (2005). Grounding emotions in human-machine conversational systems. In: Proc. of the 1st International Conference on Intelligent Technologies for Interactive Entertainment. Madonna di Campiglio, Italy, pp. 144-154.

Schatzmann, J., Georgila, K., Young, S., (2005). Quantitative evaluation of user simulation techniques for spoken dialogue systems. In: Proc. of the 6th SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal, pp. 45-54.

Schuller, B., Batliner, A., Steidl, S., Seppi, D. (2011). Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Communication. In press.

Sindlar, M., Dastani, M., Meyer, J.-J. (2010) Mental State Ascription Using Dynamic Logic. In: Proc. of the 19th European Conference on Artificial Intelligence. Lisbon, Portugal, pp. 561-566.

Sobol-Shikler, T. (2011). Automatic inference of complex affective states. Computer Speech and Language 25, pp. 45–62.

Traum, D. R. (1993). Mental state in the TRAINS-92 dialogue manager. In: Working notes of the AAAI Spring Symposium on Reasoning about Mental States: Formal Theories and Applications, pp. 143-149.

Ververidis, D., Kotropoulos, C. (2006). Emotional speech recognition: resources, features and methods. Speech Communication 48, pp. 1162–1181.

Wilks, Y., Catizone, R., Worgan, S., Turunen, M. (2011). Some background on dialogue management and conversational speech for dialogue systems. Computer Speech and Language 25(2), pp. 128-139.

Williams, J.D., Young, S. (2007). Partially observable Markov decision processes for spoken dialogue systems. Computer Speech and Language 21, pp. 393-422.

Witten, I. H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E. (2009). Reducing working memory load in spoken dialogue systems. Interacting with Computers 21(4), pp. 276-287.