18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Fusion of Sentiment Analysis and Emotion Recognition to Model the User's Emotional State

David Griol, José Manuel Molina, Jesús García-Herrero
Applied Artificial Intelligence Group
Computer Science Department
Carlos III University of Madrid - Spain
Email: {david.griol,josemanuel.molina,jesus.garciaherrero}@uc3m.es

***Abstract*—In this paper we present a framework that combines two algorithms respectively developed for Sentiment Analysis and Emotion Recognition in users spoken utterances. We propose modeling the users emotional state by means of the fusion of the outputs generated by both algorithms. This process considers the probabilities assigned to the different emotions by both algorithms. The proposed framework can be integrated as an additional module in the architecture of a spoken dialog system, using the information generated as an additional input for the dialog manager to decide the next system response.**

I. Introduction

Speech and natural language technologies allow users to communicate in a flexible and efficient manner, making possible to access applications in which traditional input interfaces cannot be used (e.g. in-car applications, access for disabled persons, etc). Also speech-based interfaces work seamlessly with small devices (e.g., smarthphones and tablets PCs) and allow users to easily invoke local applications or access remote information. For this reason, spoken dialog systems [1] are becoming a strong alternative to traditional graphical interfaces which might not be appropriate for all users and/or applications.

However, these systems are also usually designed ad-hoc for their specific domain using rule-based models and standards in which developers must specify each one of the steps to be followed by the system. This makes it difficult to adapt the resulting systems to new tasks or incorporate additional context information, as it would require modifying the hand-crafted design, which is very costly in terms of time and effort as this process cannot be automated [2], [3]. In addition, although several works emphasize the importance of taking into account context information not only to solve the tasks presented to the dialog system by the user, but also to enhance the system performance in the communication task, this information is not usually considered when designing a dialog model [4], [5].

For these reasons, the adaptation capabilities of these interfaces are frequently restricted to static choices made by the users. However, adaptation can play a much more relevant role in speech applications. The performance of a spoken dialog system also depends highly on the environmental conditions, such for example whether there are people speaking near the system or the noise generated by other devices. These systems must usually confront social, emotional and relational issues in order to enhance users satisfaction. Although emotion is receiving increasing attention from the dialog 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 [6], [7].

Emotions affect the explicit message conveyed during the interaction and is frequently mentioned in the literature as the most important factor in establishing a working alliance in dialog applications [8]. They change people voices, facial expressions, gestures, and speech speed. Emotions can also affect the actions that the user chooses to communicate with the system.

Despite its benefits, the recognition of emotions in dialog systems presents important challenges which are still unresolved. The first challenging issue is that the way a certain emotion is expressed generally depends on the speakers, their culture and environment. Another problem is that some emotional states are long-term (e.g. sadness), while others are transient and do not last for more than a few minutes. Thus, it is not trivial to select the categories being analyzed and classified by an automatic emotion recognizer. Also there is not a clear agreement about which speech features are most powerful in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch, and energy contours.

In this paper, we describe a proposal that address these important issues by developing affective dialog models for conversational systems. Our approach merges two algorithms developed within the fields of Sentiment Analysis and Emotion Recognition to respectively analyze the text transcription of the user's utterance and also consider input features extracted from the speech signal and the dialog context. The proposal is focused on recognizing negative emotions that might discourage users from employing the system again or even lead them to abort an ongoing dialog. The dialog manager of the system tailors the next system answer to the user emotional state by changing the help providing mechanisms, the confirmation strategy, and the interaction flexibility.

The remainder of the paper is as follows. In Section II

we describe the motivation of our proposal and related work. Section III describes our proposal to develop emotionally sensitive conversational interfaces. Section IV describes the application of our approach to develop a practical system providing academic information. Section V presents the results of a preliminary evaluation of this practical dialog system. Finally, Section VI presents the conclusions and suggests some future work guidelines.

II. Modeling the user emotional state

As described in the previous section, emotions can affect the explicit message conveyed during the interaction with a spoken dialog system and also the actions that the user chooses to communicate with the system. According to [9], emotions can be understood more widely as a manipulation of the range of interaction affordances available to each counterpart in a conversation. They have also been recently considered as a very important factor of influence in decision making processes. For instance, a context-aware model of emotions that can be used to design intelligent agents endowed with emotional capabilities is described in [10]. The study is complemented by also modeling personalities and mood [11].

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity [12], [13]. Opinion Mining extracts and analyzes users opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity.

Three main classification levels have been defined for SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities.

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach, and hybrid approach [12]. Machine Learning approaches apply these kinds of algorithms and uses linguistic features. Lexicon-based approaches rely on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. Hybrid approaches combine both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

Sentiment analysis is sometimes considered as a Natural Language Processing task for discovering opinions about an entity; and because there is some ambiguity about the difference between opinion, sentiment and emotion, they defined opinion as a transitional concept that reflects attitude towards an entity. The sentiment reflects feeling or emotion while emotion reflects attitude.

It was argued by Plutchik [14] that there are eight basic and prototypical emotions, which are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. Emotions Detection (ED) can be considered a SA task. SA is concerned mainly in specifying positive or negative opinions, but ED is concerned with detecting various emotions from text. As a Sentiment Analysis task, ED can be implemented using ML approach or Lexicon-based approach, but Lexicon-based approach is more frequently used.

Related to these approaches, some corpus developers prefer the number of utterances for each emotion to be almost the same in order to properly evaluate the classification accuracy. While balanced utterances are useful for controlled scientific analysis and experiments, they may reduce the validity of the data. For this reason, many other researchers prefer that the distribution of the emotions in the database reflects their real-world frequency [15], [16]. In this case, the number of neutral utterances should be the largest in the emotional speech corpus. In addition, the recorded utterances in most emotional speech databases are not produced in the conversational domain of the system [17]. Therefore, utterances may lack some naturalness since it is believed that most emotions are out comes of our response to different situations.

Very recently, other authors have developed affective dialog models which take into account both emotions and dialog acts. The dialog model proposed by [18] combined three different submodels: an emotional model describing the transitions between user emotional states during the interaction regardless of the data content, a plain dialog model describing the transitions between existing dialog states regardless of the emotions, and a combined model including the dependencies between combined dialog and emotional states. Then, the next dialog state was derived from a combination of the plain dialog model and the combined model. In our proposal, we employ statistical techniques for inferring the user's emotional state, 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 sentiment recognizers, as the intention recognizer is not linked to the dialog manager as in [18].

Van de Wal and Kowalczyk have recently present a system that automatically measures changes in the emotional state of the speaker by analyzing their voice [19]. The system was evaluated using natural non-acted human speech of 77 speakers. Chen et al. have also recently introduced an approach that combines acoustic information and emotional point information by means of SVMs, HMMs, and a soft decision strategy [20].

Bui et al. [21] based their model on POMDPs that adapt the dialog 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 dialog systems in which many emotions and dialog acts might be considered, the

authors employ decision networks to complement POMDPs. As will be described in Section III, we propose an alternative to this statistical modeling which can also be used in realistic conversational agents and evaluate it in a less emotional application domain in which emotions are produced more subtly.

Different works on audiovisual emotion recognition [22], [23], [24], [25], have shown that facial expression is a better indicator than voice for most emotions. Thus, being able to disambiguate one with the other in a multimodal system produces better results. For example, in SmartKom the results of a recognizer of emotional prosody [26] are merged with the results of a recognizer for affective facial expression [27].

In our case, we count only with the acoustic channel, so we carry out a prosody processing procedure like in SmartKom, but additionally consider other sources in order to obtain better recognition rates (as we cannot rely on other modalities). This is particularly interesting in systems in which the dialog is less flexible, where the length of the user utterances may be insufficient to enable other knowledge sources like linguistic information to be employed. That is why we propose to take into account information from the user model as well as information related to the context of the dialog that may influence the user's emotional state. This way, restricting a multimodal approach to a single modality (only voice) is not equivalent to our proposal, as we include additional sources of information that deal with the specific challenges of unimodal emotional processing.

III. Our Proposal to Develop Emotionally Sensitive Conversational Interfaces

A spoken dialog system integrates five main tasks to deal with user's spoken utterances in natural language: automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS). We propose to combine two methodologies respectively developed for emotion recognition and sentiment analysis to process the users' emotional state during the interaction, which is considered as an additional valuable input for the dialog manager to select the next system action.

A. Proposed methodology for Emotion Recognition

Our proposal to develop an emotion recognizer is based solely in acoustic and dialog 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. Our recognition method, based on the previous work described in [28], firstly takes acoustic information into account to distinguish between the emotions which are acoustically more different, and secondly dialog 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 dialog. 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 dialog information to discriminate between doubtfulness and boredom, which are more difficult to discriminate only by using phonetic cues.

This process is shown in Figure 1. As can be observed, 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 spoken dialog systems, in which a baseline algorithm which always chooses "neutral" would have a very high accuracy, which is difficult to improve by classifying the rest of emotions, that are very subtlety produced.

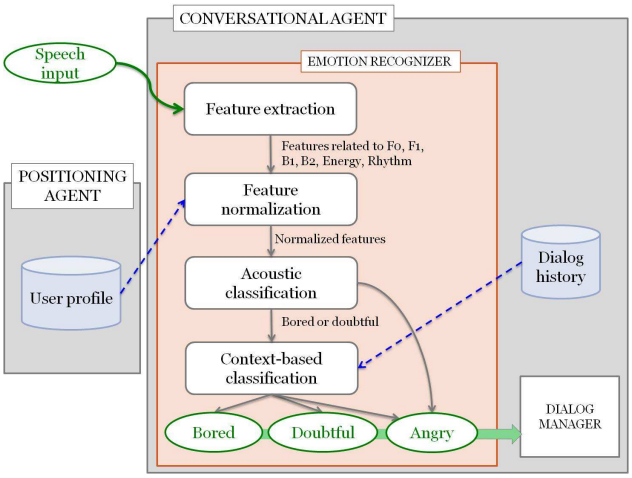

Fig. 1. Schema of the proposed emotion recognizer

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 users' voice. We extracted the most representative selection from the list of 60 features shown in Table I. The feature selection process is carried out from a corpus of dialogs on demand, so that when new dialogs 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 the Weka toolkit.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum

urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

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.

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*. The precision values obtained with the MLP are discussed in detail in [28], where we evaluated the accuracy of the initial version of this emotion recognizer. If an utterance is classified as angry, the emotional category is passed to the dialog manager of the system. If the utterance is classified as *doubtful_or_bored*, it is passed through an additional step in which it is classified according to two dialog parameters: depth and width. Dialog context is considered for emotion recognition by calculating these parameters.

Depth represents the total number of dialog turns up to a particular point of the dialog, whereas width represents the total number of extra turns needed throughout a subdialog to confirm or repeat information. This way, the emotion recognizer has information about the situations in the dialog that may lead to certain negative emotions, e.g. a very long dialog might increase the probability of boredom, whereas a dialog 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 dialog 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.

B. Proposed methodology for Sentiment Analysis

The proposed model for Sentiment Analysis aims to extend common sentiment classification of text, which is usually focused on polarity, to a higher level so that the input texts are categorized by the emotions they evoke. Thus, the main goal is to recognize a specific set of human emotions instead of only detecting whether a piece of text is negative, neutral or positive. To do this, a limited set of emotions must be selected from one of the existing emotion classifications accepted by psychologist community.

After a detailed study of the principal affective models and considering computational requirements, we have selected a modification of the Hourglass emotion representation [31]. This model is based on Plutchik's wheel of emotions, which proposes the previously described eight basic emotions contrary to Ekman's initial classification that defines only six primary affection states. Although having more categories increases analysis complexity, Plutchik's model can be reduced into four categories -as there are four pairs of opposite emotions- so that, indeed, the analysis can be considered to turn out simpler. The proposed model is based on four key components.

The Knowledge Base (KB) contains the main information sources used by the Analysis Module to extract sentiment values from words. The Analysis Module completes the words analysis. By splitting texts in sentences an tokenizing words, this module can query the Knowledge Base to extract emotional information or know whether words are modifiers or carry an associated negation. Moreover, this module identify entities in the input text and track the number of occurrences of each one of them in a similar way bag-of-words models do this using occurrences vectors.

Once the entities have been identified and words are annotated with values from the KB, the Scoring Module computes the overall relevance of the entities and assigns a weighting factor for each of the words carrying emotional information, which are also known as concepts. A weight for each of the four independent emotional categories is then computed to classify the input text.

The last stage of the model deals with knowledge learning. To do this, the Learning Module takes as input the provided analysis from users when they disagree with the results of the

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.

TABLE I
FEATURES DEFINED FOR EMOTION DETECTION FROM THE ACOUSTIC SIGNAL [29], [30], [16]

Sentiment Analysis, and computes a learning factor to modify sentiment values of involved concepts.

C. Knowledge Base

As previously described, the Knowledge Base contains the main information sources used by the Analysis Module to extract sentiment values from words. In our proposal, this information has been classified into the following categories:

- **Concepts**: A concept refers to the emotions associated to a specific pair of $(word - PoS)$, where PoS (part of Speech) denotes the grammatical function of a word inside a predicate. Only the primitive form of a word is considered and the rest of derivative words take the same set of emotional values. The different categories of words are:
  - **Nouns**: Only the singular form is considered, although they may have an irregular plural that could be harder to identify. Nouns containing prefixes and suffixes are the only exception to this rule.
  - **Adjectives**: The positive form is considered and both comparative and superlative forms are discarded.
  - **Verbs** The infinitive form is considered. Some exceptions are made for -ing forms acting as a noun (e.g., "The professor's reading about macro-economics was brilliant')'.
  - **Adverbs**: Only the positive form is considered, discarding comparative and superlative forms.
- **Modifiers**: Modifiers are denoted by an n-gram without associated sentiment states, which can increase, decrease or reverse the emotions of the associated concepts. They can be divided into two different categories:
  - **Intensity modifiers**: This category is composed by those modifiers than may increase or decrease emotions expressed by concepts (e.g., "as much" or "a bit").
  - **Negators**: These modifiers reverse the global emotion associated to a concept (e.g., "not" or "never").

The NRC[1] and SenticNet[2] emotion lexicons have been used to complete the KB. Both are publicly available semantic resources for concept-level Sentiment Analysis. A total of 12,297 concepts are currently stored in the KB.

D. Parser Module

The parsing process of a sentence generates its semantically analysis containing part-of-speech tags organized in a tree of predicates. Between the set of general-purpose libraries currently available, we have selected OpenNLP[3]. This library supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.

OpenNLP uses the Penn Treebank notation[4], which consider 36 sort of part of speech defined on the basis of their syntactic distribution rather than their semantic function. As a consequence nouns used in the function of modifiers are tagged as nouns instead of adjectives. Before parsing a text, it should be split into sentences by using the OpenNLP probabilistic *Sentence Detector*, which offers a precision of 94% and a 90% recall.

E. Emotion Classification Model

As stated before, our proposal uses an emotion representation model based on a modified version of the Hourglass model. The four independent categories that are considered for Sentiment Analysis consists of the following possible labels, described from negative maximum to positive maximum intensities, left to right:

- **Sensitivity**: [terror, fear, apprehension, neutral, annoyance, anger, rage]
- **Aptitude**: [amazement, surprise, distraction, neutral, interest, anticipation, vigilance]
- **Attention**: [grief, sadness, pensiveness, neutral, serenity, joy, ecstasy]
- **Pleasantness**: [loathing, disgust, boredom, neutral, acceptance, trust, admiration]

[1] http://www.saifmohammad.com/WebPages/lexicons.html
[2] http://sentic.net/
[3] https://opennlp.apache.org/
[4] http://www.cis.upenn.edu/ treebank/

F. Text Scoring Scheme and Adaptive Learning

Once the parsing process has finished and all the concepts, modifiers and negators have been properly tagged, it is possible to begin with the computation of the sentiment values of the text. The scoring process follows a bottom-up approach based on a fixed algorithm that relies on the Knowledge Base accuracy, a proximity based approach for modifiers, and a topic detection module to detect the most relevant topics of a text.

The way sentences are weighted is based on entities occurrences. Let w_i be the weight of a predicate and n the total number of sibling predicates that are being combined, the sentiment value of a category for weighted predicates can be defined as:

$$S_w = \frac{\sum_{i=0}^{n} w_i * s_i}{\sum_{i=0}^{n} w_i}, \quad \begin{array}{c} \forall w_i > 0 \\ \forall s_i \neq 0 \\ s_i \in [-1, +1] \\ i = [0, n] \end{array} \tag{1}$$

Our proposal also integrates an adaptive learning process for improving the Knowledge Base used for Sentiment Analysis. This process uses Eq. 2 to consider the difference between the Sentiment Analysis output proposed by the SA algorithm and the feedback provided by the user. Let U be the set of sentiments of a text corrected by the user, M be the sentiments calculated by the SA algorithm, W_{C_s} be the weight of concept C for sentiment s, and A_c be the number of accumulated adjustments of concept C. Therefore the new value of each sentiment s for a concept C is defined as:

$$C_s = C_s + \frac{(U_s - M_s) * W_{C_s}}{1 + (A_C/1000)} \tag{2}$$

IV. A CASE STUDY: THE UAH SPOKEN DIALOG SYSTEM

Universidad Al Habla (UAH - University on the Line) is a spoken dialog system that provides academic information about the Dept. of Languages and Computer Systems at the University of Granada, Spain. The information that the system provides can be classified in four main groups: subjects, professors, PhD courses and student registration [32].

A corpus of 100 dialogs was acquired with this system from student telephone calls. The total number of user turns was 422 and the recorded speech has a duration of 150 minutes. In order to develop an enhanced version of the system that includes the described algorithms for emotion recognition and Sentiment Analysis, we carried out two types of corpus annotation: intentional and emotional.

On the one hand, we estimated the user intention for 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 a user's sentence is shown below:

User Turn:
I want to know information about the subject Artificial Intelligence of Computer Science.
Semantic Representation:
(*Subject*)
Subject-Name: Artificial Intelligence
Degree: Computer Science

The labeling of the system turns was similar to that for user turns. To do so, 30 concepts were defined and grouped as task-independent concepts (e.g. *Affirmation* and *Negation*), concepts used to inform the user about the result of a specific query (e.g. *Subject* or *Lecturers*), concepts defined to require the user the attributes that are necessary for a specific query (e.g. *Subject-Name*), and concepts used for the confirmation of concepts and attributes.

On the other hand, we assigned an emotion category (neutral, doubtful, angry, or bored) to each user utterance. Nine annotators tagged the corpus twice and the final emotion for each utterance was assigned by majority voting. A detailed description of the annotation procedure and the intricacies of the calculation of inter-annotator reliability can be found in a previous study [28].

Additionally, we modified the dialog manager to process the user state information in order to reduce the impact of the user negative states and the user experience on the communication, by adapting the system responses considering user states. The dialog 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 [33], straightforward and well delimited in order not to make the user loose the focus on the task.

If the recognized emotion is doubtful and the user has changed his behavior several times during the dialog, the dialog manager changes to a system-directed initiative and generates 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 their first utterances are classified as doubtful.

In the case of anger, if the dialog 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 dialogs 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 the same 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 previous interactions and preferences.

V. Preliminary evaluation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

For comparison purposes, we have developed three versions of the UAH system: the baseline system, the ER system, and the ER+SA system. The baseline system does not carry out any adaptation to the user, using only the semantic information in the users utterances to select the next system action. The ER system integrates the proposed method for emotion recognition to use the detected emotion by means of the previously conciliation strategies.

The ER+SA system integrates the two methods developed for emotion recognition and sentiment analysis. To do this, the resulting user's emotional state is calculated by considering three cases: i) the agreement of both methods selecting the same emotion; ii) one of the two methods selects a neutral state, but the second one selects an emotion different to the neutral state with a probability higher than a given threshold; and iii) the neutral state is selected in the rest of cases.

In order to evaluate our proposal, we have recorded the interactions of 6 recruited users. Four of them recorded 30 dialogs (15 dialogs with the baseline system and 15 with the ER system), and two of them recorded 30 dialogs (15 dialogs with the baseline and 15 with the ER+SA system). Thus, a total of 180 dialogs were recorded in such a way that there were two dialogs recorded per scenario, three in the case of the five most frequent scenarios of the initial UAH corpus. An objective and a subjective evaluations were carried out.

Table II shows the results of the objective evaluation. As it can be observed, on the one hand the success rate for the ER+SA system is higher than the results obtained for the other systems. This difference showed a significance of 0.03

TABLE II
RESULTS OF THE OBJECTIVE EVALUATION OF THE SYSTEMS

Evaluation metrics	Baseline system	ER system	ER + SA system
Dialog success rate	85.0	87.0	91.0
Error correction rate	81.0	82.5	83.1
Average number of turns per dialog	12.1	11.1	10.2
Average number of actions per turn	1.8	1.5	1.5
% of different dialogs (intention and emotion)	85.0	88.0	92.0
Number of repetitions of the most seen dialog	6	3	3
Number of turns of the most seen dialog	5.5	4.6	4.5
Number of turns of the shortest dialog	4.5	4.5	4.5
Number of turns of the longest dialog	14.5	12.0	12.0

TABLE III
RESULTS OF THE SUBJECTIVE EVALUATION OF THE SYSTEMS

Questions (1 to 5 scale)	Baseline system	ER system	ER + SA system
How well did the system understand you?	4.6	4.7	4.7
How well did you understand the system messages?	3.6	3.9	3.9
Was it easy to obtain the requested information?	3.8	4.3	4.4
Was the interaction rate adequate?	3.4	4.2	4.4
If the system made errors, was it easy for you to correct them?	3.2	3.3	3.3

in a two-tailed t-test. On the other hand, although the error correction rate is also higher in absolute values in the ER+SA system, this improvement is not significant. 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. Instead, our proposal for adaptation to the user state overcomes these problems during the dialog once they are produced.

Regarding the number of dialog turns, the ER+SA system produced shorter dialogs (with a 0.00 significance value in a two-tailed t-test when compared to the number of turns of the baseline system). As shown in Table II, this general reduction appears also in the case of the longest, shortest and most seen dialogs for the enhanced system. There is also a slight reduction in the number of actions per turn for the dialogs of the ER+SA system (with a 0.00 significance value in the t-test). This might be because users have to explicitly provide and confirm more information using the baseline system, whereas the enhanced system automatically adapted the dialog to the user and the dialog history.

Regarding the percentage of different dialogs obtained, the rate was lower using the ER+SA system, due to an increment in the variability of ways in which users can provide the different data required to the enhanced system. This is consistent with the fact that the number of repetitions of the most observed dialogs is higher for the baseline system.

Table III shows the average results obtained for the subjective evaluation. As can be observed, the three 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 ER+SA system is judged to be better regarding the user observed easiness in obtaining the data required to fulfill the complete set of objectives defined in the scenario, as well as the suitability of the interaction rate during the dialog.

VI. CONCLUSIONS AND FUTURE WORK

Emotions are frequently mentioned in the literature as the most important factor in establishing a working alliance in spoken dialog systems applications. In this paper, we contribute a proposal to develop emotionally sensitive spoken conversational interfaces combining an Emotion Recognition and a Sentiment Analysis methodologies. Our proposal is focused on recognizing negative emotions that might discourage users from employing the system again or even lead them to abort an ongoing dialog. The recognized emotion is used as an additional valuable information to select and adapt the next system response.

We have also evaluated the proposed framework with the UAH spoken dialog system, implementing the prediction module between the system's natural language understanding module and dialog manager. Additionally, we have improved the dialog 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 of recruited users with the enhanced version of the system. The results show that this version of the system performs better in terms of duration of the dialogs, number of turns needed for successful dialogs, and number of confirmations and repetitions needed. Additionally, the test users judged the system to be better when it could adapt its behavior to their intentions and emotions.

As a future work we plan to annotate the emotions of the collected corpus in order to refine the adaptation strategies of the dialog manager. We also want to extend the described evaluation with a higher number of users, and also applicate the described framework to develop and evaluate additional practical dialog systems.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] R. Pieraccini, *The Voice in the Machine: Building Computers That Understand Speech*. MIT Press, 2012.

[2] T. Paek and R. Pieraccini, "Automating Spoken Dialogue Management Design using Machine Learning: An Industry Perspective," *Speech Communication*, vol. 50, pp. 716–729, 2008.

[3] J. Rouillard, "Web services and speech-based applications around VoiceXML," *Journal of Networks*, vol. 2, no. 1, pp. 27–35, 2007.

[4] S. Seneff, M. Adler, J. Glass, B. Sherry, T. Hazen, C. Wang, and T. Wu, "Exploiting Context Information in Spoken Dialogue Interaction with Mobile Devices," in *Proc. IMUx'07*, 2007, pp. 1–11.

[5] J. Ko, F. Murase, T. Mitamura, E. Nyberg, M. Tateishi, and I. Akahori, "Context-Aware Dialog Strategies for Multimodal Mobile Dialog Systems," in *Proc. of AAAI Int. Workshop on Modeling and Retrieval of Context*, 2006, pp. 7–12.

[6] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[7] M. E. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotio nrecognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[8] T. Bickmore, K. Puskar, E. Schlenk, L. Pfeifer, and S. Sereika, "Maintaining reality: Relational agents for antipsychotic medication adherence," *Interacting with Computers*, vol. 22, pp. 276–288, 2010.

[9] Y. Wilks, R. Catizone, S. Worgan, and M. Turunen, "Some background on dialogue management and conversational speech for dialogue systems," *Computer Speech and Language*, vol. 25, no. 2, pp. 128–139, 2011.

[10] G. Marreiros, R. Santos, C. Ramos, and J. Neves, "Context-Aware Emotion-Based Model for Group Decision Making," *IEEE Intelligent Systems*, vol. 25, no. 2, pp. 31–39, 2010.

[11] R. Santos, G. Marreiros, C. Ramos, J. Neves, and J. Bulas-Cruz, "Personality, Emotion, and Mood in Agent-Based Group Decision Making," *IEEE Intelligent Systems*, vol. 26, no. 6, pp. 58–66, 2011.

[12] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[13] ——, "Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[14] R. Plutchik and H. Kellerman, *Emotion: Theory, Research and Experience. Volume 1. Theories of Emotion.* Academic Press, 1980.

[15] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Getting started with susas:a speech under simulated and actual stress database," in *Proc. Eurospeech'97*, vol. 4, 1997, pp. 1743–1746.

[16] D. Morrison, R. Wang, and L. DeSilva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[17] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[18] J. Pittermann, A. Pittermann, and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *Journal of Speech Technology*, vol. 13, pp. 49–60, 2010.

[19] C. van der Wal and W. Kowalczyk, "Detecting changing emotions in human speech by machine and humans," *Journal of Applied Intelligence*, vol. 39, no. 4, pp. 675–691, 2013.

[20] L. Chen, X. Mao, P. Wei, Y. Xue, and M. Ishizuka, "Mandarin emotion recognition combining acoustic and emotional point information," *Journal of Applied Intelligence*, vol. 37, no. 4, pp. 602–612, 2012.

[21] T. Bui, M. Poel, A. Nijholt, and J. Zwiers, "A tractable hybrid DDN-POMDP approach to affective dialogue modeling for probabilistic frame-based dialogue systems," *Natural Language Engineering*, vol. 15, no. 2, pp. 273–307, 2009.

[22] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using gaussian mixture models for face and voice," in *Proc. 10th IEEE Int. Symposium on Multimedia*, 2008, pp. 250–257.

[23] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Audiovisual emotion recognition via cross-modal association in kernel space," in *Proc. ICME'11*, 2011, pp. 1–6.

[24] Z. Zeng, Y. Hu, G. Roisman, Z. Wen, Y. Fu, and T. Huang, "Audio-visual spontaneous emotion recognition," *Lecture Notes in Computer Science*, vol. 4451, pp. 72–90, 2007.

[25] L. Guan and Z. Xie, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis," *Journal of Semantic Computing*, vol. 7, no. 1, pp. 25–42, 2013.

[26] A. Batliner, R. Huber, H. Niemann, E. Noth, J. Spilker, and K. Fischer, *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000, ch. The Recognition of Emotion, pp. 122–130.

[27] W. Wahlster, *SmartKom: Foundations of Multimodal Dialogue Systems Cognitive Technologies*. Springer, 2006, ch. Dialogue Systems Go Multimodal: The SmartKom Experience, pp. 3–27.

[28] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416–433, 2008.

[29] J. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 2, pp. 151–170, 1996.

[30] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[31] E. Cambria, A. Livingstone, and A. Hussain, "The Hourglass of Emotions," *LNCS, Cognitive Behavioural Systems*, vol. 7403, pp. 144–157, 2012.

[32] Z. Callejas and R. López-Cózar, "Relations between de-facto criteria in the evaluation of a spoken dialogue system," *Speech Communication*, vol. 50, no. 8-9, pp. 646 – 665, 2008.

[33] F. Burkhardt, M. van Ballegooy, K. Engelbrecht, T. Polzehl, and J. Stegmann, "Emotion detection in dialog systems - Usecases, strategies and challenges," in *Proc. of International Conference on Affective Computing and Intelligent Interaction (ACII'09)*, 2009, pp. 1–6.