J Multimodal User Interfaces (2010) 3: 131–140
DOI 10.1007/s12193-009-0026-4



ORIGINAL PAPER

# HMM modeling of user engagement in advice-giving dialogues

**Nicole Novielli**

Received: 6 April 2009 / Accepted: 12 November 2009 / Published online: 5 December 2009


**Abstract** This research aims at defining a real-time probabilistic model of user's engagement in advice-giving dialogues. We propose an approach based on Hidden Markov Models (HMMs) to describe the differences in the dialogue pattern due to the different level of engagement experienced by the users. We train our HMM models on a corpus of natural dialogues with an Embodied Conversational Agent (ECA) in the domain of healthy-eating. The dialogues are coded in terms of Dialogue Acts associated to each system or user move. Results are quite encouraging: HMMs are a powerful formalism for describing the differences in the dialogue patterns, due to the different level of engagement of users and they can be successfully employed in real-time user's engagement detection. Though, the HMM learning process shows a lack of robustness when using low-dimensional and skewed corpora. Therefore we plan a further validation of our approach with larger corpora in the near future.

**Keywords** Real-time computational models of engagement · Embodied conversational agents · Advice-giving · Hidden Markov models · User engagement

## 1 Introduction

Our long-term goal is to build an ECA which is able to inform, persuade and engage a human interlocutor in a conversation about healthy dieting. In our simulator, the ECA plays the role of an artificial therapist and employs natural argumentation techniques to persuade the user [1]. A fundamental requirement for such an agent is the ability (i) to observe the verbal and non verbal behavior of users during the interaction, (ii) to infer the cognitive and affective ingredients of their state of mind (iii) to adapt both the dialogue strategy and interaction style accordingly [2]. This is particularly true for advice-giving dialogues in which knowledge of the user characteristics is of primary importance in building an effective persuasion strategy [1].

The main goal of advice-giving is to change, with communication, the behavior of an interlocutor by influencing her attitude (that is the system of beliefs, values, emotions that bring a person to adopt that behavior). Regardless of the application domain, a successful advice-giving strategy requires appropriate integration of two tasks: (i) provision of either general or interlocutor-tailored information about aspects of the behavior that make it more or less 'correct', and (ii) persuasion to abandon an incorrect behavior, if needed, by illustrating the negative long term consequences it entails and the positive effects of revising it.

To be effective, advice-giving cannot be the same for all interlocutors: knowledge about the user's mind should be acquired, by observing her behavior during the dialogue, in order to build a dynamic, consistent model of her mind. This model can be used for adaptation purposes and should combine both cognitive and affective ingredients of the user's state of mind. First of all, the Transactional Model by Prochaska et al. [3] suggests to adapt the dialogue strategy to the stage of change the interlocutor is in (that is, to her system of beliefs, intentions and goals). In addition, the Elaboration Likelihood Model gives a further insight on how the communication process is affected by the kind of processing the interlocutor will make of the information received. In particular, the model helps in understanding how this processing is related, at the same time, to the interlocutor's ability and

N. Novielli (✉)
Department of Informatics, University of Bari, via Orabona, 4, 70125, Bari (BA), Italy
e-mail: novielli@di.uniba.it

interest to elaborate the information received [4]: in different situations of attention and interest, peripheral or central processing channels will be followed, each focusing on a particular kind of information, with more or less emotional features. As a consequence of the two theories, knowledge about the Receiver is essential to increase her information processing ability and interest, and therefore the effectiveness of advice-giving.

In previous research, we discussed how the stage of change may be recognized dynamically during the dialogue [5]. We also proposed a method which combines linguistic and acoustic analysis for recognizing the user's 'social attitude' towards an ECA playing the role of the advice-giver [6].

In this paper, we investigate the issue of detecting the user's engagement using conversational analysis techniques. We choose the Hidden Markov Models (HMMs) [7] as a suitable formalism to represent dialogue patterns and their relations with the user's attitude. We also propose to apply this formalism in a stepwise recognition of user's engagement which enables adapting the advice-giving strategy and the system's behavior in real-time. In the study described, we propose an application of our approach to the advice-giving domain. Though, the approach itself employs dialogue pattern analysis techniques based on a domain independent framework for dialogue act annotation and hence can be easily extended to different application domains.

The paper is organized as follows: in Sect. 2 we give a brief overview of the concept of engagement in literature and provide the definition of engagement according to the advice-giving domain; Sect. 3 provides a description of the HMM-based approach, describes the corpus of dialogues used for HMM model training (Sect. 4) and gives a description of the markup language used for dialogue coding; in Sect. 5 we propose a stepwise approach for real-time detection of user engagement; conclusions and directions for future works are provided in Sect. 6.

## 2 Background

In recent years, the increasing interest of the international research community on affective computing [8] caused the flourishing of several projects aimed at recognizing the 'emotional state' of the user. Discrete sets of 'basic emotions' were recognized as well as 'emotionally-related' states (e.g. uncertainty), basic components of emotions (e.g. valence or intensity) [9–11] or personality traits [12]. These studies employed, with good recognition accuracy, a variety of acoustic and linguistic features using some classification method (discriminant analysis, naive Bayes, neural networks and others).

Rather than considering emotions, we look at two aspects of affective interaction (social attitude and level of engagement) which are presumed to be key factors for the success of the dialogue [13–15]. In particular, our previous research focused on combining linguistic and acoustic analysis of individual user's moves for detecting the social attitude of users towards an ECA [6]. In this paper we focus on conversational analysis for modeling engagement by exploiting the differences in the dialogue pattern.

Engagement is quite a fuzzy concept to which researchers attach a wide range of related but different meanings. Sidner and Lee [16, 17] talk about engagement in human-robot conversations as 'the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake'. Campbell [18] measures the involvement in spoken conversations by exploiting features which describe the synchrony between participants. For other authors, it describes 'how much a participant is interested in and attentive to a conversation' [19]. Pentland [20] defines engagement as a function of the level of involvement in the interaction. This concept is especially addressed in e-learning: here, several researchers attempted to model the attitude of students in terms of their level of initiative [21, 22] or on how much a person is being governed by the preceding interaction rather than steering the dialogue [23]. O'Brien and Toms [24] define engagement as a quality of the users' experience with technology and describe it along several dimensions including also affect and aesthetic, as well as interest and level of attention. Affect, in fact, has been demonstrated to play a fundamental role in understanding engagement: Collins [25] demonstrated that a positive emotional energy indicates a high level of engagement; Said [26] studied the emotional component of engagement of children playing video games, in terms of its correlation with the motivation for continuing to play; Smardon [27] and Olitsky [28] investigated the role of affect and the type of interactions to model students' engagement in science education.

Different definitions of engagement are meant to be coherent with the application domain and the adaptation purposes: some studies aim at implementing intelligent media switching, during human-human computer-mediated technology [19, 29]; others [21] aim at tailoring interaction to the learner's needs.

We expect the level and kind of engagement in the advice-giving task not to be the same for all users, depending on their own goals and on how useful they perceive the interaction to be: we consider users to be 'highly engaged' when the system succeeds in involving them in its persuasion attempts. A lower level of engagement, on the contrary, is attributed to users who are only interested in the information-giving task.

### 2.1 Engagement in advice-giving dialogues

In advice-giving dialogues two tasks are integrated: the provision of general or user-tailored information and the actual persuasion to abandon a problem behavior.

In a category of users, we found the typical attitude that Walton [30] calls of *examination dialogues*, in which 'one party questions another party, sometimes critically or even antagonistically, to try to find out what that party knows about something'. Examination dialogues have two main goals: the extraction of information and the testing of the reliability of this information: the testing goal may be carried out with critical argumentation aimed at judging whether the information elicited is reliable. We found this behaviour in some of our dialogues: we named the users asking several questions as *Information-Seeking* (IS), regardless of their actual intention to request information or to assess the agent's competence or the reliability of the information provided. IS users sometimes ask questions even right after the system's self introduction.

In another category (AG), users seem to be more involved in the persuasion goal of *Advice-Giving*: they show a more cooperative attitude towards the system, by providing extra-information to the agent so as to build a shared ground of knowledge about their habits, desires, beliefs etc. Also, they react to the agent's suggestions and/or attempts of persuasion by providing a constructive feedback in terms of objections, comments (either positive or negative) and follow-up questions.

Finally, we have a third category of *Not engaged* (N) users who don't show any interest in any of the two mentioned tasks (information seeking or advice-giving); they rather give a passive and barely reactive contribution to the interaction, by mainly answering the system's questions, very often with general answers (eg. 'yes' or 'no'); their dialogues are usually shorter than the others and tend to be driven by the system (that sometimes seems to struggle to protract the interaction).

Distinguishing among the three levels of engagement is relevant for our adaptation purposes: IS users might either be helped in their information seeking goal or led by the system to get involved in the advice giving task, by carefully choosing (or revising) the persuasion strategy [1]; AG users might perceive an increased satisfaction about the interaction if the agent is believable in playing the role of artificial therapist; N users represent a real challenge for the system: their attitude might be due to a lack of interest in the domain or to their being in the 'precontemplation stage' [3].

## 3 HMM modeling of engagement

This study is based on the assumption that long-term affective phenomena influence the overall behavior of users during the interaction. As a consequence, such states also impact the overall dialogue dynamics [16, 17]. This is particularly true for attitudes and social stances, which smoothly evolve during the dialogue.

Detecting long lasting features of users (such as their level of engagement) is a fundamental step towards long-term adaptation of agent's behaviour and strategy.

Our assumption is also supported by the use that researchers do of ad hoc measures for conversational analysis. Conversational turn-taking is one of the aspects of human behaviour that can be relevant for modeling social signalling [21]. In particular, Pentland [20] measures engagement by evaluating the influence that each person's pattern of speaking versus not speaking has on the other interlocutor's patterns. This is essentially a measure of who drives the conversational turn exchanges, which can be modelled as a Markov process.

Hence, we decided to model categories of users, by looking at differences in the dialogue pattern. In our previous research we argue in favor of the suitability of Markov models as a formalism for dialogue pattern description [31]: we analyze complete dialogue patterns rather than individual dialogue exchanges [32] according to our goal of predicting the user's overall final attitude. By using the formalism of HMMs, we are able to represent differences in the whole structure of the dialogues among subjects with the kinds of engagement we mentioned above.

### 3.1 Related work

HMMs find their natural and most frequent application in parsing and speech recognition [7, 33]. Their application to dialogue pattern description and recognition is more recent and this paper is a contribution in this sense.

Levin et al. [34] proposed to use this formalism in dialogue pattern modeling: system's moves are represented in states while the user's moves are associated with arcs. Their goal is to solve the problem of defining the minimal cost dialogue strategy to adopt. Stolcke et al. [35] define a discourse grammar using HMMs for Dialogue Act (DAs) prediction: they associate user moves with states in a HMM-based dialogue structure in which transitions represent the likely sequencing of user moves. Evidence about DAs are expressed in terms of their lexical and prosodic manifestations. Twitchell et al. [36] employ HMMs in classifying conversations, with no specific application reported.

The work with which our study has most in common is the analysis of collaborative distance learning dialogues by Soller [37]. This study aims at dynamically recognize when and why students have trouble in learning the new concepts they share with each other. To this purpose, specific turn sequences are identified and extracted manually from dialogue logs, to be classified as 'knowledge sharing episodes'

or 'breakdowns'. Aggregates of student's acts were associated with the five states HMMs learnt from this corpus. The overall accuracy of recognition of this study is 74%.

### 3.2 Dialogue representation

We learn our models from a corpus on natural dialogues with an ECA, collected with a Wizard of Oz (WoZ) study. Our corpus includes 30 text-based and 30 speech-based dialogues, overall 1700 adjacent pairs (system—user moves). Subjects involved were equidistributed by age, gender and background (in computer science or humanities). In our experiments, the subjects were free to interact with the ECA without any particular constraint: they could simply answer the agent's question or, instead, reply with comments, statements about their own attitudes or preferences or even ask question about the agent's life in their turn. Evaluating the effectiveness of different persuasion strategies and their impact on the user's behavior is out of the scope of the present study. We are rather interested in how the behavior of users change according to their own goals and to their level of involvement in the advice-giving task, that is we only consider the 'intention to persuade' of the agent. Hence, the Wizard's behavior stayed unvaried according to the same dialogue strategy for all the experiments. Adaptation was made only at the level of the content of the individual suggestions provided. Consistency in the Wizard's behavior was ensured through a preliminary training of the experimenter. To ensure the naturalness of the interaction, an iterative experimental design of the WoZ tool was performed. The resulting set of 86 Wizard's moves allowed us to solve the trade-off between the need of a real-time, believable reaction of the Wizard and the possibility of replying appropriately and without repetitions to the wide range of all users' moves. Moreover, the WoZ interface is designed to allow the Wizard to quickly retrieve each move during the experiment, without negatively affecting her reaction time, so as to ensure the believability of interaction. More details about the WoZ study and its design can be found in [5] and [38].

The corpus was labeled so we could capture the communicative intention of each dialogue move. Hence, we classify both system and user moves into appropriate categories of communicative acts. These categories (see Table 1) are a revision of those proposed in SWBDL-DAMSL (Switch Board Corpus—Dialogue Act Markup in Several Layers) [35]. The 86 moves the Wizard could employ (system's moves) were organized into 8 categories by considering on one hand the DAMSL classification and on the other hand the frequencies with which they had been employed in the corpus. Similar criteria were applied to define the 11 subject move categories (see Table 1).

Dialogue act sequences can easily be represented using HMMs. Formally [7, 33] an HMM can be defined as a tuple: $\langle \mathbf{S}, \mathbf{W}, \pi, \mathbf{A}, \mathbf{B} \rangle$, where

- $\mathbf{S} = \{\mathbf{s_1}, \ldots, \mathbf{s_n}\}$ is the set of states in the model;
- W is the set of observations or output symbols;
- $\pi$ are a-priori likelihoods, that is the initial state distribution: $\pi = \{\pi_\mathbf{i}\}$, $\mathbf{i} \in \mathbf{S}$;
- $\mathrm{A} = \{\mathbf{a_{ij}}\}$, $\mathbf{i}, \mathbf{j} \in \mathbf{S}$, is a matrix describing the state transition probability distribution: $\mathbf{a_{ij}} = \mathbf{P}(\mathbf{X_{t+1}} = \mathbf{s_j} | \mathbf{X_t} = \mathbf{s_i})$;
- $\mathbf{B} = \{\mathbf{B_{ijk}}\}$, $\mathbf{i}, \mathbf{j} \in \mathbf{S}$, is a matrix describing the observation symbol probability distribution: $\mathbf{b_{ij}} = \mathbf{P}(\mathbf{O_t} = \mathbf{w_k} | \mathbf{X_t} = \mathbf{s_i}, \mathbf{X_{t+1}} = \mathbf{s_j})$;

In our models *states* represent aggregates of either system's or user's moves, each with a probability to occur in that specific phase of the dialogue while the *transitions* represent the possible dialogue sequences. HMMs are learnt using the HMM Matlab Toolbox[1] from our corpus of dialogues, by representing every input dialogue as a sequence of coded dialogue moves. For example, the following dialogue excerpt:

T(S1) = Hi, my name is Valentina. I'm here to suggest you how to improve your diet. Do you like eating?
T(U,1) = Yes
T(S,2) = What did you eat at breakfast?
T(U,2) = Coffee and nothing else.
T(S,3) = Do you frequently eat this way?
T(U,3) = Yes
T(S,4) = Are you attracted by sweets?
T(U,4) = Not much. I don't eat much of them.
T(S,5) = Do you believe your diet is correct or would you like changing your eating habits?
T(U,5) = I don't believe it's correct: I tend to skip lunch, for instance.

is coded as follows: (OPENING, GENERIC-ANSWER, QUESTION, STAT-ABOUT-SELF, QUESTION, GENERIC-ANSWER, QUESTION, STAT-ABOUT-PREFERENCES, QUESTION, STAT-ABOUT-SELF).

## 4 Learning the model

Two independent raters were asked to annotate the overall attitude of each user by using the labels N, IS and AG. Both, the observed agreement (.93) and the Kappa value (.90) indicate a strong inter-rater agreement [39]. To classify the corpus by giving a final label to each dialogue, we asked the two raters to discuss the cases for which they had given different annotations. The corpus is not equally distributed (N = 28%, IS = 44%, AG = 28%), which is an undesirable circumstance when the available set of data is not particularly wide (we will show how this impact robustness of learning in Sect. 4.2).

[1] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

**Table 1** Categories of Wizard and User moves

| Speaker | Dialogue Act | Description |
|---|---|---|
| Wizard | OPENING | Initial self-introduction by the ECA |
| | QUESTION | Question about the user's eating habits or information interests |
| | OFFER-GIVE-INFO | Generic offer of help or specific information |
| | PERSUASION-SUGGEST | Persuasion attempt about dieting |
| | ENCOURAGE | Statement enhancing the user's motivation |
| | ANSWER | Provision of generic information after a user request |
| | TALK-ABOUT-SELF | Statement describing own abilities, role, skills |
| | CLOSING | Statement of dialogue conclusion |
| Subject | OPENING | Initial self-introduction by the user |
| | REQ-INFO | Information request |
| | FOLLOW-UP | Further information or justification request |
| | OBJECTION | Objection about an ECA's assertion/suggestion |
| | SOLICITATION | Request of clarification or generic request of attention |
| | STAT-ABOUT-SELF | Generic assertion or statement about own diet, beliefs, desires and behaviours |
| | STAT-PREFERENCES | Assertion about food liking or disliking |
| | GENERIC-ANSWER | Provision of generic information after an ECA's question or statement |
| | AGREE | Acknowledgement or appreciation of the ECA's advice |
| | KIND-ATTITUDE-SYSTEM | Statement displaying kind attitude towards the system (jokes, politeness, comment, question about the system) |
| | CLOSING | Statement of dialogue conclusion |

In learning HMM structures from a corpus of data, the first issue to be addressed is the number of states to use in representing the dialogue patterns. This parameter is a function of both the level of detail with which a dialogue needs to be represented, and the reproducibility of the HMM learning process, which may be represented in terms of robustness of learned structures.

The Baum-Welch algorithm [33] estimates the model parameters $\mu = (\mathbf{A}, \mathbf{B}, \pi)$ to maximize the likelihood of the given input observations, that is $\mathbf{P}(\mathbf{O}|\mu)$ (in our case, the corpus annotated according to our definition of engagement). The algorithm starts by assigning random parameters; at each iteration, the parameters are adjusted according to the maximization function. Baum-Welch is a greedy algorithm, hence it does not perform an exhaustive search of the entire solution space. On the contrary, it finds a local maximum point instead of a global one. This is a crucial issue to take into account in HMM training, especially when dealing with sparse or low-dimensional data sets.

To establish the number of states to include in our models, we tested three alternatives: 6, 8 and 10 states. The robustness analysis led us to choose the 8-state HMM model and is described in more details in [31].

### 4.1 Descriptive power of the model

Figures 1(a), (b) and (c) show, respectively, the best 8-state HMMs for N, IS and AG subjects. We denote with $\mathbf{S_i}$ and $\mathbf{U_j}$ states in which, respectively, the system (S) or the user (U) hold the initiative.

The main differences in the three models are in the Persuasion phase (S3, U3): we named this phase in each model according to the differences in the observable user categories of moves. In the N model, the users may respond to persuasion attempts with information requests, follow up questions and even with a closing move: so we named this phase *persuasion with system initiative*. IS users have the highest probability of performing a request of information and do not provide any kind of personal information (*information seeking* phase). In the AG models, users are clearly involved in an *advice-giving* phase: the probability of information requests is lower and the variety of reactions to system suggestions is wider, according to the users' goal of either enhancing the construction of a shared ground of knowledge about healthy eating, or giving a positive feedback to the ECA. Also, for this model we observe a higher likelihood of entering the persuasion phase, core of the advice-giving process, after the initial assessment of the user situation performed by the ECA through the question-answering (S2, U2). Moreover, the duration of the persuasion phase for N users is usually very short, as described by the low transition probabilities between S3 and U3. On the contrary, a long phase of either information seeking or advice-giving can be observed for, respectively, IS and AG users.

Regarding the dialogue opening (S1, U1): in the N model, U always reacts with an opening move to the self presenta-

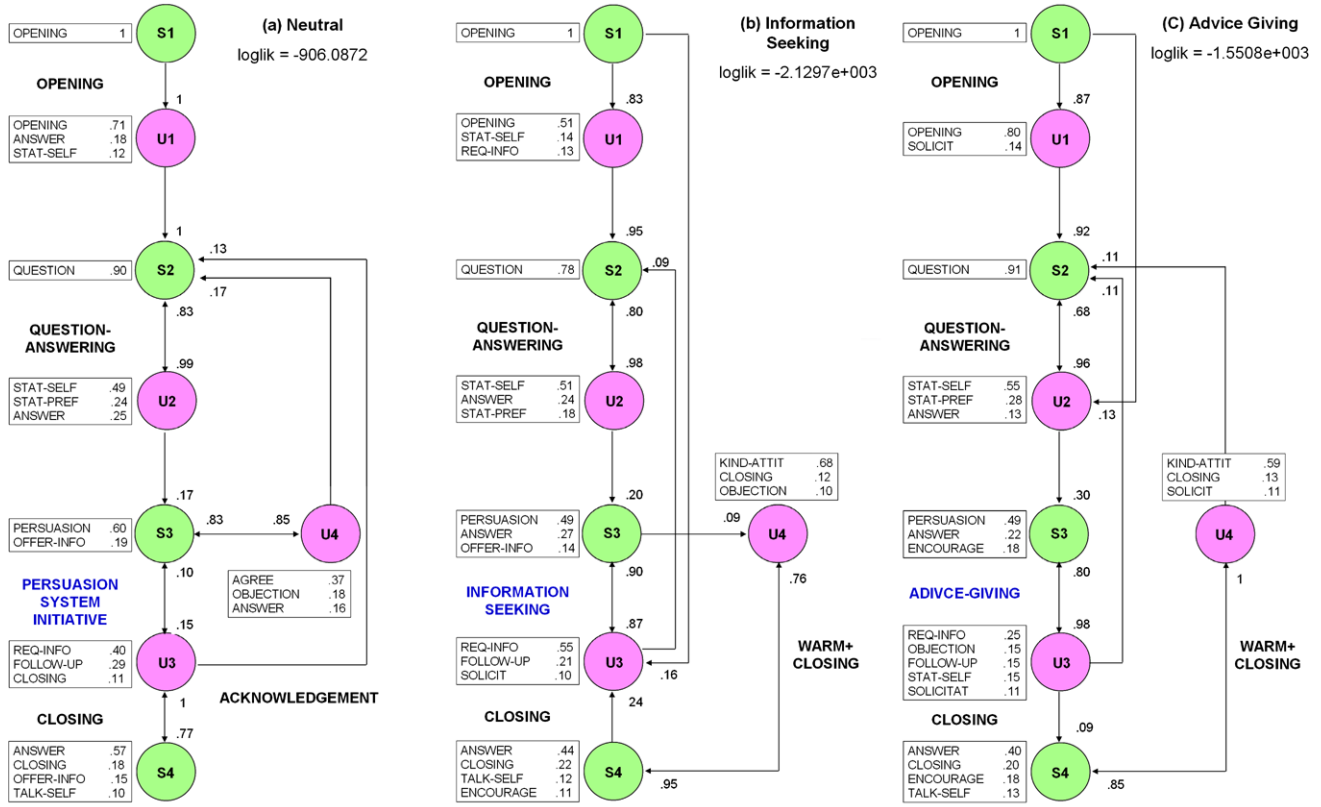


**Fig. 1** HMM for neutral (**a**), information seeking (**b**) and advice-giving (**c**) dialogues

tion of S while, in IS and AG models, there is some probability of directly entering the persuasion phase.

The question answering (S2, U2) phase proceeds basically in the same way for all dialogues, with the user replying to the ECA's questions by talking about self or providing a general answer. As hypothesized, IS and AG subjects tend to be more specific and eloquent than N ones, by producing more statements about themselves, statements about their preferences and less generic answers.

Looking at the question answering phase, what distinguishes a N from an IS or AG users is mainly its duration: we observe the highest probability of a long question answering phase for N users (see transition probability from U2 to S2) as well as the lowest probability of entering the persuasion phase (see the probability of the transition from U2 to S3) which is coherent with our definition of Not Engaged users (see Sect. 2.1). Hence, we do not consider the user's behavior during the question answering as peculiar for distinguishing among the three different levels of engagement.

Eventually, the warm phase (S4, U4) also presents some differences: in IS and AG models there is a high likelihood of observing a kind attitude, while N users mainly provide a feedback (either positive or negative) to the ECA's suggestion (acknowledgement). This can be seen as a cue of higher engagement in the interaction for IS and AG subjects. Also, contrary to IS and AG ones, in N models we notice that the probability of remaining in the persuasion phase (S3,U3) is lower than the probability of switching to the acknowledgement one; this could be seen as a proof of a low level of engagement, probably due to a lack of interest in the interaction or in the domain itself.

### 4.2 Model testing

To evaluate the classification performance of the HMM models learnt, we performed a leave-on-out validation on our annotated corpus. At every iteration $i$, the $i$-th instance of the data set is classified by choosing the model which maximizes the following: loglik = log P($i$-th case $|\mathbf{HMM_x}$), with $\mathbf{x} \in \{\mathbf{N}, \mathbf{IS}, \mathbf{AG}\}$. In other words, we choose the model that is more likely to produce the dialogue act sequence given as input. The probability is computed by using the forward-backward algorithm [33].

Table 2 shows the performance in terms of precision, recall and f1-measure. Since we are working in a supervised learning condition, we choose the most frequent label assignment (44%) as a baseline.

Results are largely above the baseline and indicate recognition performance which are comparable to analogous research in the domain [37]. In particular, we have higher performance in IS and N recognition while recognition of AG

**Table 2** Confusion matrix for N, IS and AG users

| | N | IS | AG | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| **N** | (13)**.76** | (4).24 | (0)0 | .76 | .81 | .78 |
| **IS** | (1).04 | (22)**.85** | (3).12 | .85 | .73 | .79 |
| **AG** | (2).12 | (4).24 | (11)**.65** | .65 | .69 | .67 |
| **micro** | | | | .77 | .77 | .77 |

**Table 3** Confusion matrix for N, IS and AG users

| Subjects | Distribution | Average (and variance) of $\mathbf{D(T_i, T_j)}$ | Average (and variance) of $\mathbf{D(E_i, E_j)}$ |
|---|---|---|---|
| N | .28 | .05 (.003) | .03 (.0007) |
| IS | .44 | .011 (.002) | .03 (.0003) |
| AG | .28 | .15 (.002) | .06 (.0004) |

cases seems to be the most problematic task. This is probably due to the high variety of behavior of AG users as well as the skewed distribution and restricted dimensions of our corpus, which are two of the main factors affecting the robustness of learning.

To verify this hypothesis, we conducted a further analysis in which we measured the robustness of learning by repeating the HMM learning $q = 10$ times. For each class, robustness was evaluated in terms of average differences between the $q(q-1)/2$ ($\mathbf{HMM_i}$, $\mathbf{HMM_j}$) pairs of HMMs, in the transition probabilities $\mathbf{T_i}$, $\mathbf{T_j}$ and the observation probabilities $\mathbf{E_i}$, $\mathbf{E_j}$:

- $\mathbf{D(T^i, T^j) = \sum_{h,k=1\ldots n} |T^i_{h,k} - T^j_{h,k}|/2}$
- $\mathbf{D(E^i, E^j) = \sum_{h=1\ldots n;k=1\ldots m} |E^i_{h,k} - E^j_{h,k}|/(n * m)}$

where $n = 8$ denotes the number of states and $m = 19$ the number of communicative acts used in coding. Our indices are distance metrics: the smaller the differences are the higher is the robustness of learning. Our average difference is similar to the Euclidean measure of distance between pairs of HMMs proposed in [40]. It differs from the probabilistic measure proposed in [41], in which the distance between models is measured in terms of differences in observed sequences with increasing time.

Results (see Table 3) confirm our hypothesis by showing that robustness of the method is not the same for the three classes. In spite of the high inter-rater agreement (Kappa = .90) and of the good descriptive power of the HMMs, the analysis shows a lack of robustness for AG models, due to the unequal distribution of the data-set.

The restricted amount of available data is a major cause of this phenomenon, especially when the behavior of users is extremely variable, as observed for the AG category. In fact, in the 10 repetitions of the learning experiment on AG dialogues, no clusters of similar HMMs were found, and we got the highest average differences in transitions and observations. On the contrary, in the N learning experiment (28% of cases) we had 6 similar models over 10 trained HMMs (similar likelihood values and low average differences in transitions and observations). Similarly, for the IS category, whose dialogues show a more regular structure than the AG ones, we found 7 over 10 similar models, even if the number of cases was the same as the AG.

## 5 Real-time recognition of engagement

Adaptation of the dialogue to the user's goals and preferences requires recognizing them dynamically during the interaction. In the testing method described in the previous section the whole dialogue was submitted as input.

In this section we check the ability of our classification procedure to apply a stepwise recognition method on dialogue fragments of increasing length.

Given an average number $n$ of dialogue pairs (system-user moves) considered in the training phase, we defined a 'monitoring' interval of $t$ moves and applied the recognition method to parts of the dialogue of increasing length $i * t$, with $i = 1, \ldots, n/t$. After every step, we checked whether the part of the examined dialogue fragment was recognized correctly.

Long-term affective states evolve smoothly during the interaction and drastic changes in attitude rarely occur. This is true also for the level of engagement. In particular, as already stated in Sect. 2, we aim at detecting the overall level of engagement of users, due to their own goals (i.e. cooperating in the advice-giving process vs. getting information about healthy eating) and to the perceived usefulness of the interaction. Our model of engagement is meant to be coherent with our adaptation needs: rather than classifying the user's interest or her level of attention in the next move, we want to predict the user's overall final attitude due to the her actual involvement in the task of the conversation (i.e. the advice-giving task). For this reason, we analyze

**Table 4** Stepwise recognition performance

| | | N | IS | AG | Total |
|---|---|---|---|---|---|
| a | Steadily correct recognition | .18 | .15 | .11 | .15 |
| b | Initially wrong, then correct | .65 | .46 | .50 | .52 |
| c | Steadily wrong recognition | .18 | .19 | .22 | .20 |
| d | Initially correct, then wrong | – | .04 | – | .02 |
| e | Up and down recognition | – | .15 | .17 | .11 |

complete dialogue patterns rather than individual dialogue exchanges [32]. It is reasonable to assume that the user's involvement in the domain task will probably stay unvaried and will affect her overall behavior. This assumption has been validated by our preliminary analysis of the annotated corpus of dialogues as well as the results presented in the previous section. Hence, we decided to apply a monitoring interval of $t = 4$ move pairs for engagement detection.

The findings about the descriptive power of the model (Sect. 4) and the probability distributions of our models (see Fig. 1) show how the main difference among users experiencing different levels of engagement consists in the behavior they adopt during the persuasion phase. On the contrary, the initial assessment of the situation, performed by the agent during the question-answering phase, seems to not provide any useful hint for engagement detection. Hence, we decided to not consider the user's behavior during the preliminary question-answering and we start to monitor her behavior right after the first persuasion attempt performed by the system (that is when the actual persuasion phase starts).

To test the performance of stepwise recognition, we developed a Java module which implements API for dialogue classification. Each HMM (see Fig. 1) is represented in terms of observation and transition matrices (textual format). A Java routine implementing the forward-backward algorithm is invoked whenever a new instance is given as input to the classifier. Such a module can be easily embedded in an agent-based incremental architecture [42] (e.g. in a dialogue simulator) since it has the advantage of receiving simple text as input. At every time of the interaction, automatic speech act classification is performed by a linguistic analyzer which exploit the lexical semantics of individual dialogue turns [43]. Knowledge about the system move being performed is already known.

Once again, leave one out validation was performed. We organize the results of this stepwise real-time recognition (Table 4) in five categories, according to the consequences they entail on the quality of adaptation.

The situations enabling a proper adaptation are those of 'steadily correct' (15%) and 'initially wrong then correct' (52%) recognition. In these cases a correct estimate of the user level of engagement will be performed within 4 dialogue turns after the first persuasion attempts or even earlier. This ensures a correct dialogue adaptation to the presumed user's goals since the very beginning of the persuasion phase, in the 67% of cases ($a + b$).

In a very small percentage of cases (2%) the 'initially correct, then wrong' recognition will provoke incorrect dialogue adaptation towards the end of the dialogue.

The worst cases are that of 'steadily wrong' (20%) or 'up and down recognition' (11%): here, adaptation criteria would always be wrong, or would be changed several times during the dialogue, by producing an unclear and not effective advice giving strategy.

A deeper analysis on the annotated corpus highlights that misclassified cases (especially 'steadily wrong' ones) actually are 'borderline' cases. Most part of the *c* and *e* cases, in fact, received different annotation by each of the two raters. Subjects involved in these dialogues actually show a behavior which is an hybrid behavior between IS and AG or even between N and AG ones. The 'ambiguous' behavior of these subjects often causes a misclassification of the partial dialogue input, depending on the particular peak/lack of engagement they are experiencing in a given phase of the interaction.

We individuated several possible reasons for 'borderline' behaviors, at least in our corpus of WoZ dialogues. First of all, we observed that some subjects manually annotated as N, temporarily behave like IS users because they are bored of the advice-giving process and experience the temporary goal of testing the system's ability to manage the dialogue. Hence they temporarily behave (and are misclassified) as IS, even if they don't actually have any information-seeking goal. The analysis of such cases shows that those subjects mainly rise questions which are out of the scope of the healthy-eating domain (e.g. they ask to the agent: 'Do you know who implemented you?') and change topic very frequently, which indicates that they are not actually involved in the dialogue task.

On the other side, there are also subjects who have an initially high interest in the task (either AG or IS users) but become bored, and hence not engaged, as the dialogue goes on, because they experience a mismatch between their great expectations about the application and the actual agent's competence and ability to manage the interaction.

We are currently investigating on how to integrate HMM modeling with the analysis of the content of the individ-

ual dialogue turns to disambiguate those situations, in order to enable the system to revise the dialogue strategy accordingly. Researchers working on behavioral analysis [45] propose a two layered approach combining Bayesian Networks with HMM models. This method enables integrating the HMM's ability of modeling sequences of states with the BN's ability of pre-processing multiple lower level input. In our case, HMMs learnt from dialogues about a particular category of users would be enriched by attaching to hidden states describing user moves a BN to process the evidence resulting from linguistic analysis of this move. Our expectation is that the combination of the two probability distributions of HMM and Bayesian models will improve the performance of the attitude recognition process. This approach would also allow us to realize adaptation at two levels: the overall user attitude (HMM overall prediction) and the specific signs in dialogue moves (BN prediction).

## 6 Conclusions and future work

As said in the introduction, our long-term goal is to implement an agent endowed with social and emotional skills which will play the role of artificial therapist in advice-giving dialogues. Such an agent should be able to observe the user's verbal and non-verbal behavior during the interaction, in order to infer the cognitive and affective components of her mental state. This knowledge should be formalized and used to adapt the long-term agent's behavior and the persuasion strategy accordingly [1, 2]. In this perspective, recognizing the user's attitude towards both the task of the interaction and the agent itself is a crucial issue.

In previous studies, we combined linguistic and acoustic features of the user move to dynamically build an image of her social attitude towards the agent [6, 44].

In this article, we investigated whether and how it is possible to recognize the user's level of engagement by modeling the impact of the user's attitude on the overall dialogue pattern. The main focus is on studying the suitability of the Hidden Markov Models (HMMs) as a formalism to represent differences in the dialogue model among different categories of users or long-lasting affective states, such as engagement. HMMs are recently being considered as a formalism to be applied to dialogue processing with various purposes: this work is a contribution along this perspective.

In particular we proposed a corpus-based approach to train HMMs from natural dialogues with an ECA. The models are then validated through a leave-one-out testing procedure. Real-time use of these models is implemented using a stepwise approach. In this paper we propose an application of the described approach in the advice-giving domain. Though, the method can be easily generalized according to different adaptation needs, since it is based on domain independent conversational analysis techniques and uses a domain-independent framework for dialogue act annotation.

Results are particularly encouraging and present HMMs as a promising and powerful formalism for representing differences in the structure of the interaction with subjects experiencing different levels of engagement. The main differences are observed during the persuasion phase, in which users clearly differentiate their behavior according to their engagement in the advice-giving task. Results obtained in this study also confirm our preliminary findings about attitude display in dialogue patterns [31]. In addition, the stepwise approach allows us to estimate the user's level of engagement during the interaction. In particular, the 67% of cases already receive a correct classification in the early persuasion phase. The analysis of misclassified cases suggested us to investigate the definition of a two layered approach in which we aim at combining Bayesian Networks with HMM models, so as to integrate the HMM ability of modeling sequences of states with the BN ability of pre-processing multiple level input (e.g. linguistic analysis of individual dialogue turns).

In spite of the encouraging results, though, we need to be cautious: by performing a robustness analysis with ad hoc metrics, we discovered a lack of robustness of the method that reduces the reproducibility of the learning experiment and lowers the recognition performance. We assume that this is mainly due to the dimension of our corpus and to the huge variety in the behavior of users highly engaged in the advice-giving task. Especially when combined with a low cardinality of the class, these two factors are the major causes of the reduction in the robustness of training. Future developments will involve the usage of larger corpora of data to achieve a final validation of the method, also in different application domains.

**Acknowledgements** This paper is dedicated to Fiorella de Rosis, my first mentor and advisor: without her this research would not be. I am also indebted to Berardina De Carolis for her insightful comments, helpful suggestions and continuous support. Finally, I thank Alessia Martalo' for her cooperation in the early stage of this study and Anna Rowe for kindly helping me in revising my English.

## References

1. Mazzotta I, De Rosis F, Carofiglio V (2007) PORTIA: A user-adapted persuasion system in the healthy eating domain. IEEE Intell Syst 22(6):42–51
2. Mazzotta I, Novielli N, Silvestri E, de Rosis F (2007) 'O Francesca ma che sei grulla?'. Emotions and irony in persuasion dialogues. In: Procs of the 10th conference of AI*IA—Special track on 'AI for expressive media'. AI*IA 2007: artificial intelligence and human-oriented computing, Rome. LNCS, vol 4733. Springer, Berlin, pp 602–613
3. Prochaska J, Di Clemente C, Norcross H (1992) In search of how people change: applications to addictive behavior. Am Psychol 47:1102–1114

4. Petty RE, Cacioppo JT (1986) The elaboration likelihood model of persuasion. In: Berkowitz L (ed) Advances in experimental social psychology, vol 19. Academic Press, New York, pp 123–205
5. de Rosis F, Novielli N, Carofiglio V, De Carolis B (2006) User modeling and adaptation in health promotion dialogs with an animated character. J Biomed Inf 39(5):514–531
6. de Rosis F, Batliner A, Novielli N, Steidl S (2007) 'You are soo cool Valentina!' Recognizing social attitude in speech-based dialogues with an ECA. In: Procs of ACII 2007, Lisbon
7. Charniak E (1993) Statistical language learning. MIT Press, Cambridge
8. Picard RW (2003) Affective computing: challenges. Int J Human-Comput Stud 59(12):55–64
9. Batliner A, Steidl S, Hacker C, Noth E, Niemann E (2005) Private emotions vs social interaction: towards new dimensions in research on emotions. In: Carberry S, De Rosis F (eds) Procs of the workshop on adapting the interaction style to affective factors
10. Bosma W, Andre' E (2004) Exploiting emotions to disambiguate dialogue acts. In: Proceedings of the international conference on intelligent user interfaces. Island of Madeira
11. Litman D, Forbes K, Silliman S (2003) Towards emotion prediction in spoken tutoring dialogues. In: Proceedings of HLT/NAACL
12. J Gill A, Oberlander J (2002) Taking care of the linguistic features of extraversion. In: Proceedings of the 24th annual conference of the cognitive science society
13. Bickmore T, Cassell J (2005) Social dialogue with embodied conversational agents. In: van Kuppevelt J, Dybkjaer L, Bernsen N (eds) Advances in natural, multimodal dialogue systems. Kluwer Academic, New York
14. Peters C, Pelachaud C, Mancini M, Bevacqua E (2005) Engagement capabilities for ECAs, workshop "Creating bonds with ECAs". Fourth international joint conf. on autonomous agents & multi-agent systems
15. Bickmore T, Picard R (2005) Establishing and maintaining long-term human-computer relationships. ACM Trans Comput-Human Interact
16. Sidner C, Lee C (2003) An architecture for engagement in collaborative conversations between a robot and a human. MERL Technical Report, TR2003-12
17. Sidner C, Lee C (2003) An architecture for engagement in collaborative conversations between a robot and a human. MERL Technical Report, TR2003-12
18. Campbell N (2008) Multimodal processing of discourse information; the effect of synchrony. In: Procs of the 2008 second international symposium on universal communication, pp 12–15
19. Yu C, Aoki PM, Woodruff A (2004) Detecting user engagement in everyday conversations. In: Procs of international conference on spoken language processing, pp 1329–1332
20. Pentland A (2005) Socially aware computation and communication. Computer 38(3):33–40
21. Core MG, Moore JD, Zinn C (2003) The role of initiative in tutorial dialogue. In: Procs of 10th conference of the European chapter of the association for computational linguistics, Budapest, Hungary, April 2003
22. Shah F (1997) Recognizing and responding to student plans in an intelligent tutoring system: CIRCSIM-Tutor. Ph.D. thesis, Illinois Institute of Technology
23. Linell P, Gustavsson L, Juvonen P (1988) Interactional dominance in dyadic communication: a presentation of initiative-response analysis. Linguistics 26:415–442
24. O'Brien HL, Toms EG, Kelloway EK, Kelley E Developing and evaluating a reliable measure of user engagement
25. Collins R (2004) Interaction ritual chains. Princeton University Press, Princeton
26. Said NS (2004) An engaging multimedia design model. In: Procs of the 2004 conference in interaction design and children
27. Smardon R (2004) Streetwise science: toward a theory of the code of the classroom. Mind Cult Act 11:201–223
28. Olitsky S (2007) Promoting student engagement in science: Interaction rituals and the pursuit of a community of practice. J Res Sci Teach 44(1):33–56
29. Woodruff A, Aoki PM (2004) Conversation analysis and the user experience. Digit Creat 15(4):232–238
30. Walton D (2006) Examination dialogue: an argumentation framework for critically questioning an expert opinion. J Pragmat 38:745–777
31. Martalo' A, Novielli N, de Rosis F (2008) Attitude display in dialogue patterns. In: Proceedings of AISB'08, symposium on 'affective language in human and machine'
32. Whittaker S (2003) Theories and methods in mediated communication. In: Handbook of discourse processes, LEA, Mahwah, NJ
33. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
34. Levin E, Pieraccini R, Eckert W (1998) Using Markov decision process for learning dialogue strategies. In: Proceedings of the IEEE international conference on acoustic, speech and signal processing, vol 1, pp 201–204
35. Stolcke A, Coccaro N, Bates R, Taylor P, Van Ess-Dykema C, Ries K, Shriberg E, Jurafsky D, Martin R, Meteer M (2000) Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput Linguist 26:3
36. Twitchell DP, Adkins M, Nunamaker JF, Burgoon JK (2004) Using speech act theory to model conversations for automated classification and retrieval. In: Procs of the 9th international working conference on the language-action perspective on communication modelling
37. Soller A (2004) Computational modeling and analysis of knowledge sharing in collaborative distance learning. UMUAI 14(4):351–381
38. Clarizio G, Mazzotta I, Novielli N, de Rosis F (2006) Social attitude towards a conversational character. In: Proceedings of the 15th IEEE international symposium on robot and human interactive communication. RO-MAN 2006. Hatfield, UK, pp 2–7
39. Carletta J (1996) Assessing agreement on classification tasks. The Kappa statistics. Comput Linguist 22
40. Levinson SE, Rabiner LR, Sondhi MM (1983) An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst Tech J 62(4):1035–1074
41. Juang B-H, Rabiner LR (1985) A probabilistic distance measure for Hidden Markov models. ATT Tech J 64(2):391–408
42. Carofiglio V, De Carolis B, Mazzotta I, Novielli N, Pizzutilo S (2009) Towards a socially intelligent ECA. In: Proceedings of CHItaly '09, Roma, Italy, pp 99–106
43. Novielli N, Strapparava C (2009) Towards unsupervised recognition of dialogue acts, proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: student research workshop and doctoral consortium, pp 84–89
44. Carofiglio V, de Rosis F, Novielli N (2005) Dynamic user modeling in health promotion dialogs. In: Tao J, Tan T, Picard RW (eds) Affective computing and intelligent interaction, pp 723–730
45. Carter N, Young D, Ferryman J (2006) A combined Bayesian Markovian approach for behaviour recognition. In: Procs of the 18th international conference on pattern recognition