Attitude Display in Dialogue Patterns

Alessia Martalò[1]**, Nicole Novielli**[1] **and Fiorella de Rosis**[1]

Abstract. We investigate how affective factors influence dialogue patterns and whether this effect may be described and recognized by HMMs. Our goal is to analyse the possibility of using this formalism to classify users' behavior for adaptation purposes. We present some preliminary results of an ongoing research and propose a discussion of open problems.

1 INTRODUCTION

Advice-giving is aimed at attempting to change, with communication, the behavior of an interlocutor in a given domain, by influencing his or her attitude (the system of beliefs, values, emotions that bring the person to adopt that behavior). Irrespectively of the application domain, this goal requires appropriate integration of two tasks: provision of general or interlocutor-specific information about aspects of the behavior that make it more or less 'correct', and persuasion to abandon a problem behavior, if needed, by illustrating negative long term consequences it entails and positive effects of revising it.

To be effective, advice-giving cannot be the same to all interlocutors. According to the Transactional Model, it should be adapted, first of all, to the *stage* at which the interlocutors may be located, in the process of passing from a 'problem' to a 'more correct' behaviour [1]: that is, to their beliefs, intentions and goals. In addition, the effect of the communication process will be conditioned by the *kind of processing* the interlocutor will make of information received. In this case, the Elaboration Likelihood Model helps in understanding how this processing is related, at the same time, to the Receiver's ability and interest to elaborate it [2]. 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. The consequence of the two theories is that, in advice-giving dialogues, knowledge of the Receivers is essential to increase their information processing ability and interest, and therefore the effectiveness of advice-giving.

In previous papers, we discussed how the stage of change may be recognized and updated dynamically during the dialogue [3]. We also discussed how the user's 'social attitude' towards the advice-giver -in our case, an Embodied Conversational Agent (ECA)- could be recognized with a combination of language and speech [4]. In both cases, the unit of analysis was the individual user move: results were propagated in a dynamic probabilistic model, to progressively build an approximate image of the user.

In this article, we wish to discuss whether the user attitude reflects into the overall dialogue pattern rather than into the linguistic or the acoustic features of individual moves. We consider Hidden Markov Models (HMMs) as a candidate formalism to represent dialogue patterns and their relations with the user attitude. We also propose to apply this formalism in a stepwise recognition of this attitude that enables adapting the advice-giving strategy and the system behavior.

The paper is organized as follows. In Section 2, we clarify the aspects of the user attitude we intend to recognize and model: in particular, 'engagement'. In Section 3, after briefly introducing the kind of data on which our models were built, we describe how we applied HMMs to learn dialogue pattern descriptions for various user categories. In Section 4 we test the descriptive power of HMMs when trying to model the differences in dialogue dynamics between two classes of users, defined according to their background. Model testing is discussed in Section 5, while the topic of engagement recognition is dealt with in Section 6, before a brief analysis of related work (Section 7) and some final considerations about the limits of this ongoing study (Section 8).

2 WHICH ATTITUDE ASPECTS

Knowledge of the user characteristics is of primary importance when trying to build an effective persuasion strategy: this knowledge may be acquired by observing the users' behavior during the dialogue to build a dynamic, consistent model of their mind. This model can be used for adaptation purposes and should combine both affective and cognitive ingredients. 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 [5].

2.1 Social attitude

With the term *social attitude* we intend *"the pleasant, contented, intimate feeling that occurs during positive interactions with friends, family, colleagues and romantic partners... [and] ... can be conceptualized as... a type of relational experience, and a dimension that underlines many positive experiences."* [6]. Researchers proposed a large variety of markers of social presence related to verbal and nonverbal behaviour [7,8,9]. By grounding on these theories, in a previous research we proposed a method to recognize social attitude in dialogues with an ECA by combining linguistic and acoustic analysis of individual user moves [4].

2.2 User engagement in advice-giving

Engagement is a quite fuzzy concept, as it emerges from analysis of the literature, to which researchers attach a wide range of related but different meanings. Sidner and Lee [10] 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*". To other authors, it describes "*how much a participant is interested in and attentive to a conversation*" [11]. Pentland [12] engagement is a function of the level of

[1] Dept. of Informatics, Univ. of Bari, Via Orabona 4, 70126 Bari, Italy
Email: **alessiamartalo@libero.it**
{**novielli,derosis@di.uniba.it**}

involvement in the interaction, a concept especially addressed in the e-learning domain. Here, several researchers attempted to model the attitude of students in terms of their level of initiative [13, 14] or of how much a person is being governed by the preceding interaction rather than steering the dialogue [15].

Different definitions of engagement are meant to be coherent with application domain and adaptation purposes: some studies aim at implementing intelligent media switching, during human-human computer-mediated technology [11,16]; others [13] aim at tailoring interaction to the learner's needs. Our long-term goal is to implement a dialogue simulator which is able to inform and persuade a human interlocutor in a conversation about healthy dieting. We expect the level and kind of engagement in the system goals not being 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

2.3 Attitude display and conversational analysis

This study is based on the assumption that affective phenomena influence the dialogue dynamics [10]. For this reason, we decided to model categories of users, by looking at differences in the dialogue pattern. Our assumption is supported by the usage that researchers do of *ad hoc* measures for conversational analysis, by taking into account several dimensions (linguistic and prosodic features) and units of analyis (phonemes, words, phrases, entire dialogues). Conversational turn-taking is one of the aspects of human behaviour that can be relevant for modeling social signalling [13]. Pentland [12] 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. In a previous research, we dynamically estimated the probability value of social attitude [3] by looking at linguistic and acoustic evidences at the single user move level [4], to adapt the style of the next agent move. Detecting long lasting features of users (such as their level of engagement) can be seen as a further step towards long-term adaptation of agent's behaviour and strategy. Also, we believe that such features have a long-term impact on the overall behaviour of users. For this reason, we analyse complete dialogue patterns rather than individual *dialogue exchanges* [17]: rather than classifying the next user move, we want to predict their overall final attitude. By using the formalism of HMMs, we aim at representing differences in the whole structure of the dialogues among subjects with the kinds of engagement we mentioned above.

3 MODEL LEARNING

After becoming a very popular method in language parsing and speech recognition [18,19], Hidden Markov Models are, more recently, being considered as a formalism to be applied to dialogue processing with various purposes: to describe and classify dialogue patterns in various situations and to recognize the category to which new dialogues probably belong. This new application domain requires careful critical analysis, to understand the conditions under which successful studies can be performed. This paper is a contribution in this direction.

3.1 Corpus description

Our corpus includes 30 text-based and 30 speech-based dialogues with an ECA, collected with a Wizard of Oz (WoZ) study: overall, 1700 adjacent pairs (system – user moves). Subjects involved were equidistributed by age, gender and background (in computer science or humanities).

3.2 Corpus labelling

The corpus was labelled so as to classify both system and user moves into appropriate categories of communicative acts. These categories were a revision of those proposed in SWBDL-DAMSL (Switch Board Corpus - Dialogue Act Markup in Several Layers) [20]. The 86 moves the Wizard could employ (system moves) were organized into 8 categories (Table 1) by considering on one hand the DAMSL classification and on the other hand the frequencies with which they had been employed in the corpus.

Tag	Description
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 aimed at enhancing the user's motivation
ANSWER	provision of generic information after a user request
TALK-ABOUT-SELF	statement describing own abilities, role and skills
CLOSING	statement of dialogue conclusion

Table 1: Categories of Wizard moves

Similar criteria were applied to define the 11 subject move categories (Table 2):

Tag	Description
OPENING	initial self-introduction by the user
REQ-INFO	information request
FOLLOW-UP-MORE-DETAILS	further information or justification request
OBJECTION	objection about an ECA's assertion or 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	acknowledgment or appreciation of the ECA's advice
KIND-ATTITUDE-SYSTEM	statement displaying kind attitude towards the system, in the form of joke, polite sentence, comment or question about the system
CLOSING	statement of dialogue conclusion

Table 2: Categories of subject moves

3.3 Dialogue representation

Formally [18,19], an HMM can be defined as a tuple: < S, W, π, A, B>, where

- S = $\{s_1, \dots s_n\}$ is the set of states in the model,
- W is the set of observations or output symbols,

- π are a-priori-likelihoods, that is the initial state distributions: $\pi = \{\pi_i\}, i \in S$;
- $A = \{a_{ij}\}$, $i, j \in S$, is a matrix describing the state transition probability distribution: $a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i)$;
- $B = \{b_{ijk}\}$, $i, j \in S$, $w_k \in W$, is a matrix describing the observation symbol probability distribution: $b_{ijk} = P(O_t = w_k \mid X_t = s_i, X_{t+1} = s_j)$.

In our models:

- *States* represent aggregates of system or user moves, each with a probability to occur in that phase of the dialogue.
- *Transitions* represent dialogue sequences: ideally, from a system move to a user move type and vice versa, each with a probability to occur (although in principle, user-user move or system-system move transitions may occur).

HMMs are learnt from a corpus of dialogues by representing the input as a sequence of coded dialogue moves. For example, the following dialogue:

T(S,1)= 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 jump 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).

3.4 Setting the number of states in the model

In learning HMM structures from a corpus of data, one has first of all to establish the number of states with which to represent dialogue patterns [2]. This is a function of at least two factors: (i) the *level of detail* with which a dialogue needs to be represented, and (ii) the *reproducibility of the HMM* learning process, which may be represented in terms of *robustness of learned* structures.

The Baum-Welch algorithm adjusts the model parameters $\mu = (A, B, \pi)$ to maximize the likelihood of the input observations, that is $P(O|\mu)$. The algorithm starts with random parameters, and, at each iteration, adjusts them according to the maximization function. This algorithm is, in fact, very similar to the Expectation-Maximization (EM) algorithm and, like this, is *greedy*. That is, it does not explore the whole solution space (not exhaustive search) and can find a local maximum point instead than a global one: this is the reason why, as we will see later on, repeated applications of the algorithm to the same dataset may produce different results.

To establish the number of states with which to represent our models, we tested three alternatives: 6, 8 and 10 states. For each condition, we repeated q times the learning experiment with the same corpus of data, in identical conditions. Robustness of learning was evaluated from the following indices:

- *loglik values*: groups of HMMs with similar logliks were considered to be similar;
- *average differences* between the q*(q-1)/2 (HMM^i, HMM^j) pairs of HMMs, in the transition probabilities T^i, T^j and the observation probabilities E^i, E^j:

$$D(T^i,T^j) = \Sigma_{h,\,k = 1,\ldots,n} \; | T^i{}_{h,k} - T^j{}_{h,k}| / n^2$$

$$D(E^i,E^j) = \Sigma_{h = 1,\ldots,n;\; k = 1,\ldots,m} \; | E^i{}_{h,k} - E^j{}_{h,k}| / n * m$$

where n denotes the number of states (6, 8 or 10) and m the number of communicative acts used in coding (19). Differences are computed after aligning the states in the two models. Our average difference is similar to the Euclidean measure of distance between pairs of HMMs that was proposed in [21]. It differs from the probabilistic measure proposed in [22], in which the distance between models is measured in terms of differences in observed sequences with increasing time.

States	Average (and variance) of D(Ti,Tj)	Average (and variance) of D(Ei,Ej)
6	.18 (.005)	.043 (.0003)
8	.041 (.001)	.014 (.00009)
10	.055 (.0005)	.022 (.00011)

Table 3: Comparison of HMMs with different n. of states

Table 3 shows the results of this analysis on the corpus of 30 text-based dialogues, after repeating q=10 times the learning experiment. HMMs with 8 states are the most 'robust' as they show the minimum average difference in transitions and observations. At the same time, they are quite easy to interpret, as they do not include 'spurious' states assembling system and user moves at the same time. HMMs with 10 states provide a more detailed dialogue description but are much less robust. We therefore selected the 8-state option for our following experiments. Robustness of learning measures how reproducible the learning process is. For instance, in the ten repetitions of the experiment on text-based dialogues, 7 over 10 HMMs had very similar loglikelihood values and low average differences of transitions and observations; they could therefore be considered as 'similar' and 'good' models. This result may be interpreted by saying that the probability of finding a 'good' model by repeating the learning experiment is .70. Although this is not a high value, we could notice, in our subsequent experiments, that other categories of dialogues were still less robust. In general, with the increasing complexity of dialogues in a category (as in the case of those collected with speech-based interaction), model learning becomes less robust. As we will see, this lack of robustness affects considerably the quality of our results.

4 DESCRIPTIVE POWER OF THE MODEL

To test the descriptive power of HMMs learnt from our corpus of data, we considered a user feature about which we had acquired some knowledge in a previous study. In that study we analysed the relationship between user background (in computer science -CS- or in humanities -HUM-) and 'social attitude' towards the ECA [4]. As we mentioned in the Introduction, the method of analysis employed was, in that context, language and speech processing of individual moves. Results of that study proved that users with a background in humanities displayed a different behavior in dialogues: they tended to have a 'warmer' social attitude towards the ECA, that was displayed with a familiar language, by addressing personal questions to the agent, by talking about self etc.

[2] We used HMM Matlab Toolbox: http://www.ai.mit.edu/~murphyk/Software/HMM/hmm.html

Figures 1 (a) and (b) show, respectively, the best 8-states HMMs for CS and HUM subjects. States Si correspond to aggregates of system moves: in 1a, an OPENING is associated with S1 with probability 1; a QUESTION to S2 with probability .88; a PERSUASION (p=.57), an OFFER-INFO (p=.20) or an ENCOURAGE (p=.14) with S3, etc. Interpretation of states Uj, to which user moves are associated, can be observed from the figure. Transitions between states in models 1a and 1b have a common core pattern, although with different probabilities: the path (S1, U1, S2, U2, S3, U3), the way back (U3, S2) and the direct link (S1, U3). Other transitions differ. Dissimilarities can be found also in the probability distributions of communicative acts associated with the phases of dialogue opening (S1, U1), question answering (S2, U2), system persuasion (S3, U3) and of a warm phase (S4,U4), in which the user displays a kind attitude towards the system in various forms. The following are the main differences between the models, in these phases:

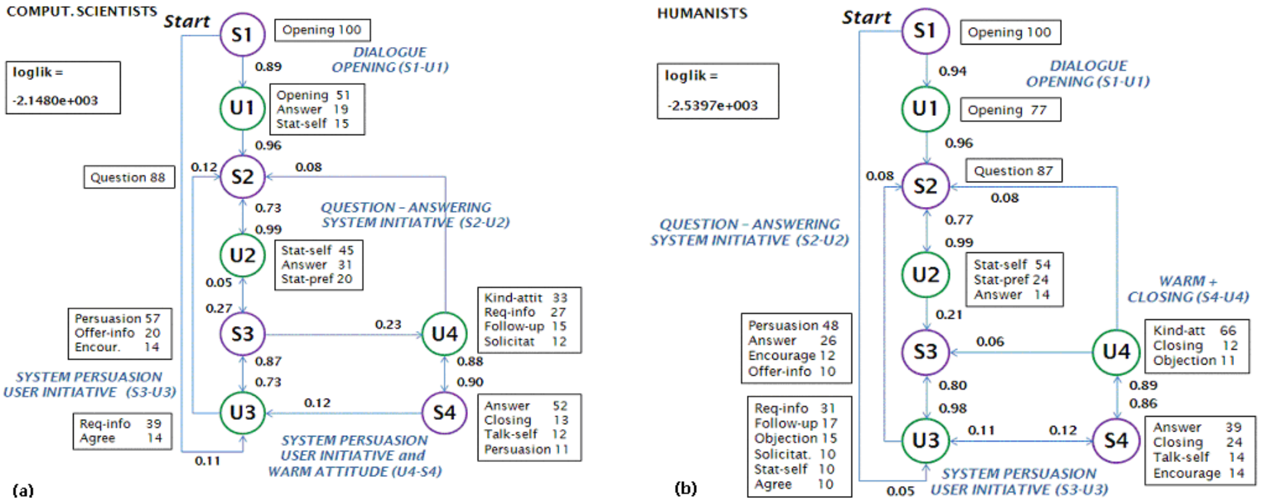

Fig 1: HMMs for subjects with a background in computer science (a) and humanities (b)

- *Question answering* (S2, U2)**:** the only difference, in this case, is that HUM subjects tend to be more specific and eloquent than CS ones, by producing more "statements about self", "statements about preferences" and less "generic answers".
- *Persuasion* (S3, U3)**:** in CS models, users may respond to persuasion attempts with information requests or declarations of consensus. They may enter, as well, in the warm phase (S3, U4 link). In the HUM model, after a persuasion attempt by S, U may stay in the persuasion phase (U3) by making further information requests, objections, solicitations or statements about self. In both models, question answering and persuasion may alternate (link U3, S2) in a more varied dialogue.
- *Warm phase* (S4, U4): although this phase exists in both models, communicative acts employed by U, once again differ: the 'objection' move of HUM is substituted by a 'solicitation' in CS's model. In this model, the state S4 may contain persuasion moves as well: hence, the whole phase may be called "system persuasion -user initiative- and warm attitude". The likelihood to produce a 'kind attitude' move (in U4) is, in the HUM model, twice than in the CS model.

These differences in the subjects behavior confirm the findings of our previous studies, by describing them in terms of dialogue structure rather than of individual moves' content.

5 MODEL TESTING

Before applying HMMs to reason about the differences among dialogues produced by different types of users, we needed to test the ability of learnt HMMs to classify correctly new cases.

5.1 Method

We describe the method to classify new cases in two classes: this can be easily extended to the case of p>2 classes. Given a corpus of dialogues classified in two sub-corpora (S-C1 and S-C2), of dimensions n and m, collected from two different categories of users (C1 and C2) according to a given target feature - for example interaction mode (text-based vs speech-based) or background (CS vs HUM) -:

a. Train two HMMs, respectively from n-1 cases from S-C1 and the m cases in S-C2, and call them HMM1 and HMM2.

b. Test the n-th case on HMM1 and HMM2 with the forward-backward algorithm, to compute the loglikelihoods:

$$loglik1 = \log P(\text{n-th case} \mid HMM1)$$
$$loglik2 = \log P(\text{n-th case} \mid HMM2)$$

c. Select the maximum between loglik1 and loglik2 to attribute the n-th case to C1 or C2.

d. Check the correctness of the classification (matching with an 'external reference').

e. Repeat from a. to d. by varying the training set according to the leave-one-case out approach.

f. Repeat from a. to e. (m-1 cases in S-C2 and n cases in S-C1).

g. Compute the recognition accuracy for HMM1 and HMM2 in terms of confusion matrix, precision and recall.

5.2. Recognizing the user background

By applying HMM analysis to users' background, we aimed at verifying whether any difference between the two typologies of users could be found, as well, in the dialogue patterns. Information about this feature was collected with a questionnaire during the WoZ study. Here, it was taken as the ‘external reference’ in the testing phase. Table 4 shows the results of this analysis: a CS dialogue is recognized correctly in 77 % of cases, while a HUM dialogue is recognized correctly in 57% of cases. HUM dialogues tend to be confused with CS ones more frequently than the inverse.

	CS HMMs	HUM HMMs	Total
CS dialogues	(23) .77	(7) .23	30
HUM dialogues	(13) .43	(17) .57	30
Total	36	24	
Recall	.77	.57	
Precision	.64	.71	

Table 4: Confusion matrix for CS vs HUM dialogues

5.3. Stepwise recognition

Adaptation of the dialogue to the user goals and preferences requires recognizing them dynamically during the dialogue. In the testing method we described in the previous Section, on the contrary, the whole dialogue was submitted to testing. We therefore wanted to check the ability of our classification procedure to apply a stepwise recognition method on subsections of dialogue 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, ... n/t. After every step, we checked whether the part of the dialogue examined was recognized correctly. What we expected from this analysis was to find an increase of recognition accuracy with the increasing monitoring time.

To check the validity of the method, once again we applied stepwise recognition to the distinction between CS and HUM dialogues, with a monitoring interval of t=5 pairs. The results we got were less positive than our expectation. In Table 5, results of stepwise recognition are classified in five categories, according to the consequences they entail on the quality of adaptation. The worst cases are that of ‘steadily wrong’ (22%) or ‘up and down recognition’ (15%): here, adaptation criteria would be always wrong, or would be changed several times during the dialogue, by producing an unclear and not effective advice giving strategy.

	CS	HUM	Total
Steadily correct recognition	14 (47%)	9 (30%)	23 (38%)
Initially wrong, then correct	2 (7%)	8 (27%)	10 (17%)
Steadily wrong recognition	7 (23%)	6 (20%)	13 (22%)
Initially correct, then wrong	1 (3%)	4 (13%)	5 (8 %)
Up and down recognition	6 (20%)	3 (10%)	9 (15 %)

Table 5: Stepwise recognition for CS vs HUM dialogues

In the cases in the ‘initially correct, then wrong’ category (8%) adaptation would become incorrect towards the end of the dialogue while, in the cases in the ‘initially wrong, then correct’ category (17%), the system might adopt correct adaptation criteria only towards the end of the dialogue. The only situation enabling a proper adaptation is that of ‘steadily correct recognition’ (38%). Notice that in HUM dialogues (which, as we said, are longer and more complex) it takes more time to the system to recognize properly the user category. We attributed this poor stepwise recognition ability to the limited robustness of both the learning and the testing procedure, due to the reduced dimension of our corpus. To test this hypothesis, we repeated the stepwise testing on the same dialogue with the same learned HMM, and applied ‘majority agreement’ as a criterion for recognizing the background at every step. We did this little check with 6 dialogues and 5 repeated tests, but found some improvement of results only in some of them .

6 RECOGNIZING ENGAGEMENT

In this section we present a possible application of the method described in Sections 3 to 5. In particular, we aim at testing whether HMMs can be employed to represent differences in the dialogue pattern of users which show different goals and levels of involvement in the advice-giving task.

In advice-giving dialogues two tasks are integrated: provision of specific information about aspects of the behavior that make it more or less ‘correct’, and persuasion to abandon a problem behavior. In a category of users, we found the typical attitude that Walton [23] 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 are shown to have two goals: the extraction of information and the testing of the reliability of this information: this testing goal may be carried out with critical argumentation used to judge whether the information elicited is reliable. We found this behaviour in some of our dialogues: we therefore named “information-seeking” (IS) the users asking several questions, either for requesting information or challenging the application, sometimes 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 toward 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 adaptation: IS users might be either helped in their information seeking goal or leaded by the system to get involved also in the advice giving task, by carefully choosing (or revising) the persuasion strategy [25]; 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’ [1].

6.1 Corpus annotation

Two independent raters were asked to annotate the overall attitude of every user by using the labels N, IS and AG. The percentage of agreement (.93) and the Kappa value (.90) indicate a strong interrater agreement [24]. To classify the corpus by giving a final label to every dialogue, we asked the two raters to discuss about the cases for which they had given different annotations. The resulting distribution of the corpus is skewed, which is an undesirable circumstance when the available set of data is not particularly wide: we will show how robustness of learning decreases if compared with the previous classification attempts, were the corpus was equally distributed.

6.2 HMM training and robustness

To evaluate how suitable are HMMs to model user engagement we repeated the robustness analysis described in 3.4, with 8-state models. Results (in Tab. 6) show that the robustness of the method is not the same for the three classes.

Subjects	Distribution (%)	Average (and variance) of D(Ti,Tj)	Average (and variance) of D(Ei,Ej)
N	.44	.05 (.003)	.03 (.0007)
IS	.28	.011 (.002)	.03 (.0003)
AG	.28	.15 (.002)	.06 (.0004)

Table 6: Robustness evaluation

In spite of the high interrater-agreement and of the good descriptive power of the HMMs, the analysis shows a lack of robustness, especially for AG models, due to the unequal distribution of the dataset. The restricted amount of available data is a major cause of this phenomenon, especially when the behaviour of users is extremely variable, as observed for the AG category. In fact, in the 10 repetitions of the learning experiment on these 28 dialogues, no groups of similar HMMs were found, and we got the highest average differences in transitions and observations. On the contrary, in the N learning experiment (44 cases) we had 6 similar models over 10 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 for the AG. This confirms our findings about the CS vs HUM classification experiments, by adding an extra insight due to the unequal distribution of the corpus.

	N	IS	AG	Total
N	(13) .76	(4) .24	(0) 0	17
IS	(1) .04	(22) .85	(3) .12	26
AG	(2) .12	(4) .24	(11) .65	17
Total	16	30	14	
Recall	.76	.85	.65	
Precision	.81	.73	.69	

Table 7: Confusion matrix for N, IS and AG

The results for the leave one case out validation (tab. 7), performed on the three classes of engagement, confirm, once again, that the higher is the variety of behaviour among users in a given class, the worse is the recognition performance (lowest value for both precision and recall, for AG users).

6.3 Descriptive power: classification of engagement

In spite of the issues highlighted in par. 6.2, HMMs still seem to meet our expectation about their ability of distinguishing among the three levels of engagement we wish to recognize. Figures 2 (a), (b) and (c) show, respectively, the best 8-states HMMs for N, IS and AG subjects.

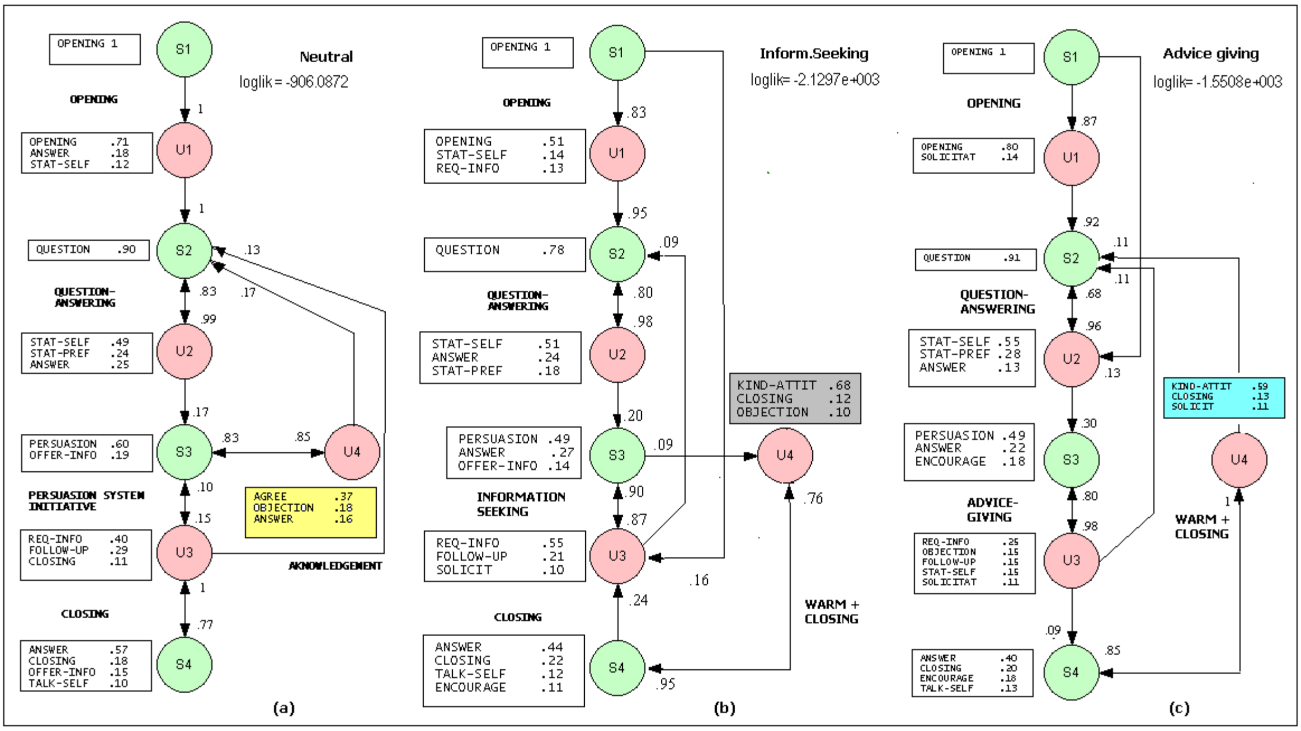

Fig 2: HMM for neutral (a), information seeking (b) and advice-giving (c) dialogues

The following are the main differences between these models:

- *Dialogue opening* (S1, U1): in the N model, U always reacts with an opening move to the self presentation of S while, in IS and AG models, there is some probability of directly entering the persuasion phase;
- *Question answering* (S2, U2): as hypothesized, IS and AG subjects tend to be more specific and eloquent than N ones, by producing more "statements about self", "statements about their preferences" and less "generic answers";
- *Persuasion* (S3, U3)**:** we named the persuasion phases according to the differences in the observable user categories of moves. In the N models, 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 AG models, users are 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. The likelihood of entering the persuasion phase, core of the advice-giving process, after question answering, is higher in these models;
- *Warm phase* (S4, U4): 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 low level of engagement, probably due to a lack of interest in the interaction or in the domain itself.

The comments we just provided depict HMMs as a powerful formalism for differentiating among various categories of users. The lack of robustness of the method, though, suggests us to be cautious: we decided to describe the three best models (with best loglikelihoods) but the limited amount of training data and the huge variety in users' behavior, especially for the AG category, affect the reproducibility of the learning experiment and cause poor recognition performance. Our findings about the descriptive power of HMMs should therefore be validated by further investigation on larger corpora.

7 RELATED WORK

HMMs find their more natural and more frequent application domain in parsing and speech recognition problems. Their application to dialogue pattern description and recognition is more recent. Levin et al [26] were among the first authors to use this formalism in dialogue modeling. In their proposal, system moves are represented in states while user moves are associated with arcs. The costs of different strategies are measured in terms of distance to the achievement of the application goal (information collection in an air travel information system), and the optimal strategy is the minimal cost one. In [20], user moves are associated with states in a HMM-based dialogue structure, transitions represent the likely sequencing of user moves, and evidence about dialogue acts are their lexical and prosodic manifestations. Twitchell et al [27] proposed to use HMMs in classifying conversations, with no specific application reported.

The work with which our study has more in common is the analysis of collaborative distance learning dialogues in [28]. The aim, in this case, was to dynamically recognize when and why students have trouble in learning the new concepts they share with each other. To this purpose, specific turn sequences were 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 of HMMs learnt from this corpus. The overall accuracy of recognizing effective vs ineffective interactions was 74 %.

8 CONCLUSIONS AND OPEN PROBLEMS

There are, in our view, some new aspects in our study, from both the methodological and the result points of view. In previous studies, we combined linguistic and acoustic features of the user move to dynamically build an image of his/her social attitude towards the agent. In this article, we investigated whether and how it is possible to model the impact of user's attitude on the overall dialogue pattern. In particular, we aim at: (i) studying the suitability of the HMMs as a formalism to represent differences in the dialogue model among different categories of users, as differentiated by either stable user features (such as background) or long-lasting affective states (such as attitudes); (ii) highlighting the importance of evaluating the robustness of the trained HMMs before using them, to avoid the risk of building unreproducible models; (iii) proposing the usage of robustness metrics for assessing the role played by the size of the dataset used and how this affects the performance of recognition tasks.

We first tested the descriptive power of HMMs with a pilot experiment: two models were trained from our corpus of data to classify users on the basis of their background, a stable and objective user feature whose role and impact on interaction dynamics have been widely investigated in our previous research. We then extended the method to the recognition of three classes of users, who showed different levels of engagement in the advice-giving task.

Results present HMMs as a suitable and powerful formalism for representing differences in the structure of the interaction of subjects belonging to different categories (Sections 4 and 6). Still, we have to be cautious: all the models described in this paper are those who showed the best training likelihood. By using *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 assumed that this is mainly due to the dimension of our corpus. Also, the complexity of dialogues and the huge variety in the behaviour of users of certain classes (e.g., people with background in humanities and AG users, especially when 'naturally' interacting via speech) play an important role in this sense. Especially when combined with a low cardinality of the class, these two factors are, in our opinion, the major causes of the reduction in the robustness of training. Future developments will involve the usage of a larger corpus of data, to achieve a final validation of the method.

Another open problem is how to use these models to dynamically recognize user attitudes (such as engagement), for long-term adaptation of the agent's behaviour. In Section 5.3, we tested a possible stepwise approach to simulate the usage of HMMs during the interaction: the idea is to define/revise the ECA's dialogue strategy according to the predicted overall level of engagement, to prevent involved users to be unsatisfied or to try to enhance involvement of those users who show a lack of interest in the advice-giving task. Results of the stepwise simulation are not encouraging, again probably because of the limited amount of data we used for training. Poor recognition performances are obtained especially when dialogues belonging to the same class have particularly complex dynamics and there is high variability among them (e.g. HUM users). The results of the experiment show that a proper adaptation could be possible for only 38% of cases. In all the other cases, results of the recognition would lead the system to an unclear and uneffective persuasion strategy: whether this adaptation approach would be successful and would produce a significant increase of the level of engagement in the users is still not clear and should be further investigated. Researchers working on behavioral analysis [29] proposes 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 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 allow us to realize adaptation at two levels: the overall user attitude (HMM overall prediction) and the specific signs in dialogue moves (BN prediction).

REFERENCES

[1] J. Prochaska, C. Di Clemente and H. Norcross. In search of how people change: applications to addictive behavior. *Americal Psychologist*, 47, 1102-1114, 1992.

[2] R. E. Petty and J.T. Cacioppo. The Elaboration Likelihood Model of Persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*. New York: Academic Press, 19, pp. 123-205, (1986)

[3] F. de Rosis, N. Novielli, V. Carofiglio and B. De Carolis. User modeling and adaptation in health promotion dialogs with an animated character. *Journal of Biomedical Informatics,* 39 (5), 514-531 (2006)

[4] F. de Rosis, A. Batliner, N. Novielli and S. Steidl. 'You are soo cool Valentina!' Recognizing social attitude in speech-based dialogues with an ECA. In: *Procs of ACII 2007*, Lisbon (2007)

[5] T. Bickmore and J. Cassell. Social Dialogue with Embodied Conversational Agents, in: J. van Kuppevelt, L. Dybkjaer, & N. Bernsen (Eds.), *Advances in Natural, Multimodal Dialogue Systems*. New York: Kluwer Academic (2005)

[6] P.A. Andersen and L.K. Guerrero. Handbook of Communication and Emotions. Research, theory, applications and contexts. Academic Press, New York, (1998)

[7] Polhemus, L., Shih, L-F and Swan, K., 2001. Virtual interactivity: the representation of social presence in an on line discussion. *Annual Meeting of the American Educational Research Association*.

[8] K. Swan. Immediacy, social presence and asynchronous discussion, in: J. Bourne and J. C. Moore (Eds.): *Elements of quality online education*. Vol. 3, Nedham, MA. Sloan Center For Online Education (2002)

[9] J.N. Bailenson, , K.R. Swinth,, C.L Hoyt, S. Persky, A. Dimov, and J. Blascovich. The independent and interactive effects of embodied agents appearance and behavior on self-report, cognitive and behavioral markers of copresence in Immersive Virtual Environments. *PRESENCE*. 14, 4, 379-393 (2005)

[10] C. Sidner and C. Lee. An architecture for engagement in collaborative conversations between a robot and a human. *MERL Technical Report*, TR2003-12 (2003)

[11] C. Yu, P. M. Aoki and A. Woodruff. Detecting User Engagement in everyday conversations. In *Procs of International Conference on Spoken Language Processing.* 1329-1332 (2004)

[12] A. Pentland. Socially Aware Computation and Communication. *Computer*, 38, 3, 33-40 (2005)

[13] M. G. Core, J. D. Moore, and C. Zinn, 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)

[14] F. Shah. Recognizing and Responding to Student Plans in an Intelligent Tutoring System: CIRCSIM-Tutor. Ph.D. thesis, Illinois Institute of Technology (1997)

[15] P. Linell, L. Gustavsson, and P. Juvonen. Interactional dominance in dyadic communication: a presentation of initiative-response analysis. *Linguistics*, 26:415–442 (1988)

[16] A. Woodruff, P. M. Aoki. Conversation analysis and the user experience. *Digital Creativity*, 15 (4): 232-238 (2004)

[17] S. Whittaker. Theories and Methods in Mediated Communication, in *Handbook of Discourse Processes*, LEA, Mahwah, NJ (2003)

[18] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. In: *Procs of the IEEE*, 77,2, 257-286 (1989)

[19] E. Charniak, *Statistical language learning*. The MIT Press (1993)

[20] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26, 3 (2000)

[21] S. E. Levinson, L. R. Rabiner and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *B:S:T:J.*, 62, 4, 1035-1074 (1983)

[22] B-H Juang and L R Rabiner. A probabilistic distance measure for Hidden Markov Models. *AT&T Technical Journal*. 64, 2, 391-408, (1985)

[23] D. Walton: Examination dialogue: an argumentation framework for critically questioning an expert opinion. *Journal of Pragmatics,* 38,745-777 (2006)

[24] J. Carletta: Assessing agreement on classification tasks. The Kappa statistics. *Computational Linguistics*, 22 (1996)

[25] I. Mazzotta, F. De Rosis and V. Carofiglio. PORTIA: A user-adapted persuasion system in the healthy eating domain. *IEEE Intelligent Systems*, in press.

[26] E. Levin, R. Pieraccini and W Eckert. Using Markov decision process for learning dialogue strategies. Proceedings of the *IEEE International Conference on Acoustic, speech and signal processing,* 1, 201-204 (1998)

[27] D. P. Twitchell, M. Adkins, J. F. Nunamaker and J. K. Burgoon. 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* (2004)

[28] A. Soller. Computational modeling and analysis of knowledge sharing in collaborative distance learning. *UMUAI*, 14, 4, 351-381, (2004)

[29] N. Carter, D. Young and J. Ferryman. A Combined Bayesian Markovian Approach for Behaviour Recognition. In Proceedings of the 18th International Conference on Pattern Recognition, (2006)