# Adapting sentiment analysis to face-to-face human-agent interactions: from the detection to the evaluation issues

Caroline Langlet
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
46 Rue Barrault
75013 Paris, France
Email:caroline.langlet@telecom-paristech.fr

Chloé Clavel
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
46 Rue Barrault
75013 Paris, France
Email: chloe.clavel@telecom-paristech.fr

***Abstract*—This paper introduces a sentiment analysis method suitable to the human-agent and face-to-face interactions. We present the positioning of our system and its evaluation protocol according to the existing sentiment analysis literature and detail how the proposed system integrates the human-agent interaction issues. Finally, we provide an in-depth analysis of the results obtained by the evaluation, opening the discussion on the different difficulties and the remaining challenges of sentiment analysis in human-agent interactions.**

***Index Terms*—other-repetition; engagement; alignment; emotional stance; embodied conversational agent**

## I. Introduction

In the research field of the embodied conversational agents (ECA), the detection of sentiment-related phenomena is a challenging task contributing to the improvement of the human-agent interactions. Various works provide systems able to detect non-verbal clues of the user's affective reactions (facial or bodily expressions and acoustic features [1]). The verbal content is more and more integrated but still partially exploited. The few linguistic-based systems for the detection of the user's sentiments concern more avatars and visualization than interaction issues [2], [3]. The only work integrating a module for the detection of user's sentiments in verbal content uses methods which are not different from sentiment analysis and opinion mining [4].

The research field of sentiment analysis (SA) provides a set of methods for the detection of sentiment-related phenomena in texts. Several works have provided machine learning methods for the classification of words, sentences or texts [5], [6], [7], [8]. In order to identify and analyse more precisely the expressions of sentiment, some works adopt fine-grained methods dealing with the syntactic structure of sentences. These representations can be applied in either rule-based [9], [10] or machine learning [11] or hybrid algorithms [12] (combining machine learning approaches and in-depth sentence analysis).

These methods show good results and interesting prospects. However, they are designed as text-mining applications. Their goal is to provide information about customers or social network users. Consequently, the extracted information and the used methods are not always suited to human-agent interaction applications. From the ECA's point of view, the goal of understanding the user's sentiments is to inform an socio-emotional agent and to help him to build communicational strategies.

Designing a system able to detect the sentiment-related phenomena in the context of a face-to-face interaction requires tackling different issues [13]. First of all, we have to define the relevant user's sentiment expressions according to the agent's communicational goals: the detection system has to select, among the expressions of sentiment-related phenomena uttered by the user, the ones which are relevant for producing the agent's reactions and building long-term relationships. Secondly, the system needs to be designed for the processing of a face-to-face conversational speech. So, it must process the user's utterances not in a fragmentary way but by considering the dialogue context. Finally, the system has to be evaluated according to the requirements of both natural language processing (NLP) and virtual agent research. These two research fields provide evaluation protocols suitable to the properties of the systems they design. The evaluation of a NLP system embedded in the virtual agent system needs to check the recall and the precision but also the semantic information provided to the agent. In order to deal with these issues, we provided a rule-based system designed according to the human-agent interaction requirements. In [14], we detailed all the semantic and syntactic rules that it embeds. The present paper highlights how our method differs from SA methods regarding not only the design but also the evaluation protocol, and contributes to human-agent interactions by dealing with the conversational speech and the agent's communicational goals.

First, we make a connection between ECA issues and the linguistic models used in SA in order to select one sentiment-related phenomenon interesting for modelling the social rela-

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

tionships: the user's likings that are given by the expressions of user's likes and dislikes in the verbal content (Section II). Second, we present the positioning of our system and of the evaluation protocol according to the existing SA literature and detail how the proposed system integrates the human-agent interaction issues (Section III). Finally, we provide an in-depth analysis of the evaluation results, opening the discussion on the different difficulties and the remaining challenges of sentiment analysis in human-agent interactions (Section IV).

## II. Delimiting User's Relevant Sentiment-Related Phenomena

The system introduced in the next section aims to detect sentiment-related phenomena by using a rule-based and symbolic method. This system adapts a sentiment analysis approach by considering the liking dimension. The underlying sentiment theoretical model has been chosen according to: (i) its ability to provide a complex framework enabling the selection of relevant sentiment-related phenomena (ii) its description of sentiment expression. This section details the connection that we made between the ECA theoretical background and the linguistic models used in SA. After a description of the liking dimension in the ECA systems, we examine the ability of the sentiment theoretical models to deal with expressions referring to user's liking.

### A. The Liking Dimension in Existing ECA Systems

In the research field of ECA, several studies – as [15] or [16] – aim to design applications involving social relationships between the ECA and the human user. In order to model relationships, these studies use several dimensions. The liking dimension is one of the most widely used. The definition of this concept is frequently grounded on the Heider's *Balance Theory* [17] which is concerned with "the way relations among persons involving some impersonal entity are cognitively experienced by the individual" [18]. It considers the relations between a person P, which is the focus of the analysis, another person O and an entity X, which can be either an event, a process or a physical object. These relations can be modelled by a triangle where the vertices are P, O and X (Figure 1). The edges are the liking relations between them : P's liking toward O, P's liking toward X, and O's liking toward X. For Heider, "a balanced state exists if all three relations are positive in all respects or if two are negative and one positive" [17]. Thus, if P likes X, O likes X and P likes O, then the state between P, O and X is balanced. Similarly, if both P and O dislikes X, and P likes O, then the state is balanced. Otherwise, the stated is unbalanced, that it can be a source of strain for P. As the theory assumes that to reduce the tension, people tends to keep balanced states, some social agent systems define scenarios where user (P)'s liking toward the agent (O) is determined by the user and the agent's feeling toward a impersonal entity (X). If the agent and the user share the same feeling toward the impersonal entity (X), then the user's liking toward this agent is positive, otherwise it is negative.

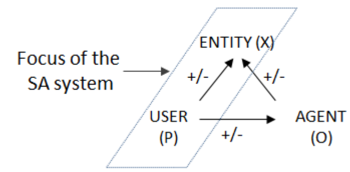


Fig. 1. Heider's Balance Theory

The detection of user's likings toward different impersonal entities will contribute to help the agent to balance the states and reduce tensions. Thus, in such a context, the analysis of user's verbal content has a key role as a major source of information of user's likings.

### B. Linguistic and Psychological Models in Sentiment Analysis

Sentiment-related phenomena belong to a very large category which can comprise various kind of expressions. In the research field of sentiment analysis, several words are used for referring to the targeted phenomenon: sentiment, opinion, judgement, appreciation, textual affect. Each of these words refer to phenomena which present both specific features and overlaps [13]. As it provides conceptual definition determining the candidate expressions for the detection, the choice of a relevant theoretical model is an important step. Furthermore, according to its complexity, it can offer detailed and useful classification for building consistent outputs.

In the research field of SA, several studies focus on the objective-subjective opposition for classifying texts, sentences and words. This distinction is also completed by the valence opposition (positive-negative) which is generally grounded on the Osgood's dimensional model [19] or the Private State Theory [20]. In order to classify more precisely words and sentences, several studies refer to psychological models. For example, in [21] and [22], the authors use the classification of emotions provided by [23], whereas in [24], they refer to the OCC (Ortony, Clore & Collin) model [25]. As an in-depth approach, those psychological models offer a detailed description of sentiment-related phenomena. However, they are rather designed for describing psychological processes than verbal expressions.

From the human-agent point of view, a valence classification of the user's sentence is a useful first step. However, the valence axis can not be used alone as it cannot model the complexity of sentiment phenomena. In particular, the valence axis does not delimit the user's expressions of likes and dislikes, which correspond to a subset of sentiment-related phenomena as described in the next subsection. If the psychological models can handle the issues, they do not provide linguistic descriptions of sentiment-related phenomena.

### C. An Appropriate Phenomenon: the User's Likes and Dislikes

The model proposed by Martin and White [26] is increasingly used in several studies [9], [27], [28]. As a systemic linguistic model, it aims to adapt the appraisal model to verbal issues. In this way, it provides a complex framework of attitudes. Three classes of attitude are defined: the affects, which are concerned with emotional reactions; the judgements, which appraise people's behaviours; the appreciations, which relate to evaluations toward semiotic or natural phenomena. As appraisal processes, the attitudes rely on three elements: the source, the person evaluating or experiencing the attitude, the entity which is evaluated or which triggers this affect, and a linguistic clue expressing the evaluation. Our computational model adapts this theory to the agent's communicational goals by selecting, among all the expressions of attitude, those which can refer – explicitly or implicitly – to a user's like or dislike. This sub-set of attitudinal expressions is established from the three original categories described by [26]. For example, the sentences "This painting makes me sad" and "This painting is a master-work", which can refer, in a different way, to the source's liking, belong respectively to the affect and the appreciation categories.

## III. Providing a Detection System and an Evaluation Protocol Designed for the Conversational Context

From the ECA point of view, a SA system has to provide a fine-grained analysis for distinguishing several expressions in a same sentence and identifying the source and the target for each sentence. Furthermore, the modularity is needed in order to make our system able to process verbal content by dealing with the progression of the conversation. Finally, it needs to be evaluated by a protocol considering the conditions of the analysis process. In [14], we introduced both the linguistic rules embedded by the first version of our system and the evaluation protocol. The present section details how both the system and the evaluation have been designed to deal with the issue of analysing human-agent conversations.

### *A. Detection Systems and Evaluation Protocols in SA*

*1) From Subjectivity Classification to Fine-Grained Analysis:* A large number of studies in SA use machine learning algorithms for the classification of words, sentences or texts. Various algorithms have been used: *Support Vector Machine* [5] [6], *Naives Bayes classifier* [7], *conditional random fields* [8]. Whether supervised or not, these approaches show interesting results. In the conversational context, they can be used, such as [4], to attribute a polarity to a speech turn. Nevertheless, they are designed for processing large corpora and do not consider expressions in an individual way. Thus, they are not well-suited to a conversational context. Moreover, these methods do not properly manage the detection of the target and the source.

In order to prevent these issues, some works provide fine-grained approaches dealing with the sentence structure either in rule-based, machine learning or hybrid algorithms. By taking into account the syntactic and semantic structure of utterances, such approaches can deal with the compositionality

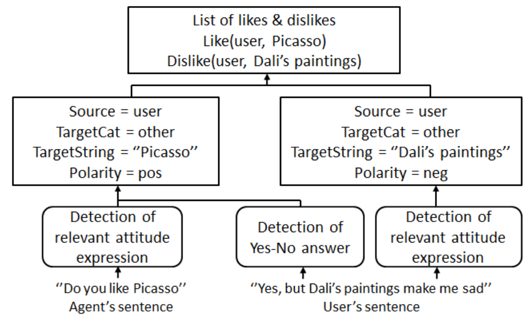


Fig. 2. Process overview

principle [1] and improve the calculation of the polarity. While in [22] and [10], the authors define semantic rules grounded on the hierarchic relations within the sentences, in [11], they train a recursive neural network on a sentiment treebank. In both cases, they can deal with polarity reversal and propagation. These approaches also appear as well-suited for the identification of the source and target. In [29] and [12], the authors design a hybrid method, jointly using extraction patterns and conditional random fields to maximize the identification of the source and target.

*2) Evaluation Protocols Adapted to Text-Mining Issues:* The SA community provides evaluation protocols which mostly rely on annotated datasets. Three types of annotations can thus be found in the literature. The first type consists in using web resources which already include internet users annotations. For example, in [5], the authors use the website rottentomatoes.com for extracting film reviews provided by internet users and the scores attributed to each film as a reference annotation. These resources are a cost-saving and time-efficient solution, however the data and the annotations that they provide match with specific application goals.

The second type corresponds to manual annotation campaigns calling expert external annotators. In this case, the annotation scheme models the sentiment-related phenomenon on which the computational system is grounded on. If the annotation scheme are more or less detailed, their goal is to provide a finer classification of the sentiment-related phenomena than the web resources provides. These endeavours usually entrust the annotation task to three or four expert annotators, i.e. linguists [9].

The third type uses crowd sourcing plateforms for building annotated datasets. For example, the Amazon Mechanical Turk (AMT) plateform has been used for various tasks of language annotation – evaluation of machine translation [30], affect recognition [31], or dictionary validation [32]. They observe a high agreement of non-expert raters with the gold standards.

[1]This principle defines the meaning of a sentence as built by the meaning of its constituents

### *B. A System Focusing on the Adjacency Pairs*

In order to provide a fine-grained analysis of the user's likes and dislikes, we choose a rule-based method relying on symbolic representations of sentence. With this approach, we deal with the conversational context by modelling the agent's utterances for detecting the user's likes and dislikes. As shown in a previous work [33], a large part of the attitudes expressed in a dialogue context are built in a collaborative way: they can be triggered by the agent's solicitations or grounded on its verbal content. For example, some user's utterances may express an expression of attitude without containing any lexical clue. In the following example, *Agent:"Do you like this painting?" – User: "Yes, I do"*, no attitude can be detected by focusing on the analysis of the user's sentence.

For this reason, we design a system able to process jointly the agent and the user's speech turns within adjacency pairs (AP) [34] (see Figure 2). Concerning the agent's speech turn, the system detects attitudes – referring to a like or a dislike – with either an affirmative or interrogative form and whose source can be either the agent or the user. This analysis is required as the agent's sentences are scripted and not automatically generated in the used ECA platform [35]: no semantic features can be obtained as internal value. Concerning the user's speech turn, we aim to detect attitudes referring to likes and dislikes whose source is the user. Moreover, in the user's speech turn, the system is able to detect expressions of agreement or disagreement which can validate or invalidate the expressions of attitude detected in the agent's sentence.

For both the agent and the user's speech turns, the expressions of like-dislike are detected by using a bottom-up and rule-based process, which launches successively different analysis levels: a lexical level, a chunk level and then a sentence level [14]. These three stages comprise syntactic and semantic rules used for assigning a polarity to the chunks and sentences detected and for identifying the source and target. These rules are grounded on the syntactic form of like-dislike expressions and deal with the compositionality principle: by first detecting lexical clues of like-dislike expressions (using [36] and by then parsing grammatical structures, they enable the identification of dependency relations and semantic roles of the constituents. Relying on this information, the system assigns polarity to the head constituent and identifies, at the final stage, the sources and the targets of the expressions. The rules have been designed according to both the literature and the study of the Semaine corpus.

### *C. An Evaluation Protocol Suitable to a SA System Embedded in an ECA Plateform*

The annotation protocol needs to be defined according to our application context. The few studies providing systems able to detect sentiment-related phenomena in user's verbal content does not consider the dialogue context. As a first work for handling this issue, our system deals with agent's utterances within APs and without considering the multimodal clues and a larger dialogue context. Our evaluation protocol takes into account these processing conditions by integrating a system-oriented annotation task. We supply to the annotators the same input as the system: transcribed adjacency pairs without the conversational context and the multimodal context.

Moreover, unlike SA evaluation protocols, our goal is not to obtain manual annotations for training a machine learning algorithm but for verifying that our system makes human-like interpretations. In this way, we do not need to know how experts analyse likes and dislikes in the verbal content, but how ordinary locutors recognize and interprete them. Thus, we choose to entrust the annotation task to non-experts labellers. The crowd-sourcing website Amazon Mechanical Turk (AMT) was used to carry out the annotation campaign and to quickly obtain a large number of annotators. Recent works have shown the reliability of the annotations obtained via AMT [30], [31], [32], and observe a high agreement of non-expert raters with the gold standards.

The AMT workers have been selected according to their approval rate – greater than or equal to 99% – and to the number of task approved – greater than or equal to 10000. As the annotation is done by non-expert annotators, we designed a simplified and intuitive annotation process: for each pair of speech turns, the annotator have to answer to a set of questions. This questionnaire has been built to determine whether the annotator is able to deduce like or dislike from the user's speech, (see Figure 3). In order to facilitate the annotation and to make the interpretation of each sentence as spontaneous as possible, the question have been designed without linguistic technical word. In this way, the task is more functional for the annotator and it is easier for him/her to put his/herself in the place of the hearer.

This annotation protocol has been applied to a dataset comprising 60 sub-sets of 10 pairs of sentences, extracted from the Semaine Corpus [37]. Each subset of the corpus is randomly assigned to one annotator, and the order in which the AP are presented to each annotator is also randomly defined. A training phase is previously subjected to each annotator in order to familiarise him/her to the annotation principles. Finally, 240 AMT workers have participated to the annotation campaign (4 for each subset). In [14], we detail all the inter-agreement measures. With regard to the presence of a like-dislike expression in a pair, the Fleiss' Kappa [38] varies depending on the sub-sets: 40% of the subsets has a kappa score comprised between 0.40 and 0.60. For the polarity, 52% of the sub-corpus has a percentage of agreement greater than 75%. Regarding the target, 61% of the sub-corpus has a percentage of agreement greater than 50%.

## IV. Discussing Disagreement and Errors

The system has been evaluated according to a reference made of 496 pairs showing a consensus between the 4 annotators. The agreement is substantial for the detection of the presence of an expression ($k = 0.61$), the number of user's expressions ($\alpha = 0.67$ [2]) and the polarity ($k = 0.844$).

[2] We use the Cronbach's alpha [39]

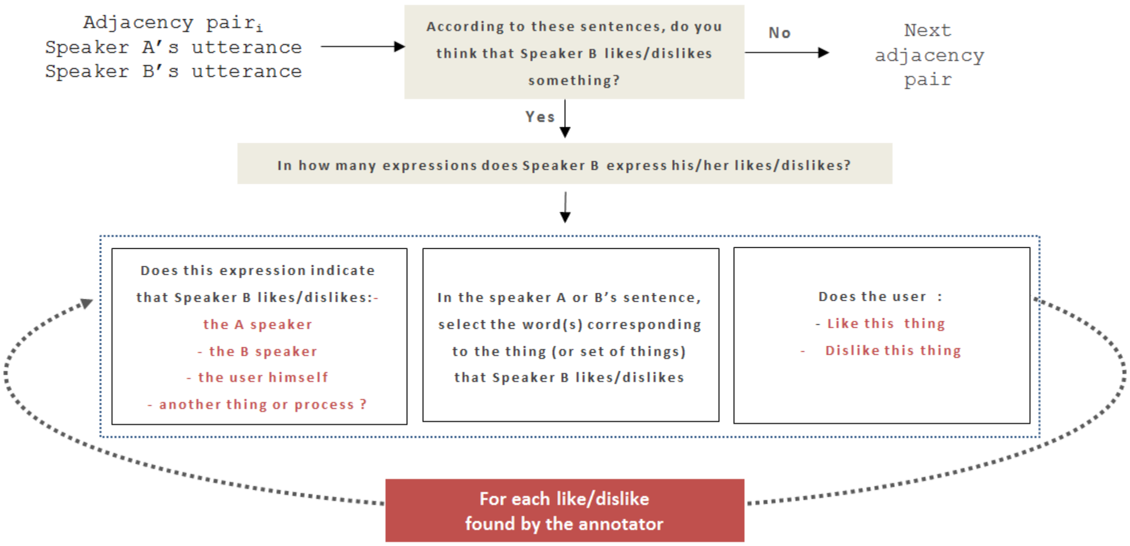


Fig. 3. Annotation process

Regarding the target type, we obtain a percentage of agreement at 53%. This section presents an in-depth analysis of the disagreement between the system and the reference. It aims to highlight the complexity of sentiment expressions in a conversational context and to provide tracks for improving the system in a future work.

### A. *Explicit vs Implicit Expressions of Like or Dislike*

*1) Implicit Expressions as a Source of Disagreement:* Concerning the presence of a like-dislike expression, 4% of the pairs (22 pairs) are annotated by the system as containing one expression, while the referred annotation does not indicate the presence of any like/dislike expression. Mostly, this disagreement concerns sentences comprising an attitudinal word but which only implicitly refer to a like or a dislike (16 over 22).

We define as implicit, the user's expressions of like or dislike which refers to a negative or positive evaluation toward an object, without using a predicate referring to the process of like-dislike (either verbs such as "love", "enjoy", "like", "hate" or nominal predicate such as "feel love", "having affection"). In our context, this kind of expression can be uttered either in the user's sentence or in the agent's sentence with a user's validation or invalidation. Our system considers them as clue of user's like or dislike. However, the reference is not so systematic. For example, regarding the user's sentence in the pair, Agent: "so are you having a nice day today" – User: "it's ok I did some stats this morning so that's always good", our system considers the positive evaluation "that's always good" as a clue of the user's like toward "stats". However, the reference does not consider the pair as conveying a user's expression of like.

The mechanism is almost the same for the user's validations-invalidations of implicit expressions conveyed by the agent's sentences. For example, in the following pair, Agent: "well then maybe it's not a good idea to get angry very often" – User: "yeah that's true", the system interprets the negative appreciation "it's not a good idea to get angry very often" – expressed by the agent and validated by the user – as a clue of the user's dislike regarding the targeted behaviour "get angry very often".

This disagreement regarding the implicit expressions also occurs between the human annotators. We examine, in the subsets with a Kleiss' Kappa under the average, the data where no consensus can be found. In most cases, the disagreement involves the user's sentences containing an expression of attitude which refers to an evaluation but not directly to a like or dislike process. It seems that, with this kind of sentences, while some annotators act as our system and consider the positive or negative feeling expressed as a clue of the user's like or dislike, other annotators have a stricter interpretation of the phenomenon. They only gather, in the category of like-dislike expressions, the explicit sentences. Disagreement between annotators also frequently occurs when this kind of expression is uttered in the agent's sentence with a user's validation or invalidation. For example, in the following pair, Agent: "excellent, excellent, I love happy things in life" – User: "me too", while two annotators have rated the pair as containing a user's expression of like, the two others do not consider this expression as a relevant clue of a user's like.

This disagreement between, human annotators, on the one hand, and the system and the reference, on the other hand, highlights the difficulty to strictly position the borderline of like-dislike expressions.

*2) Agreement on the Explicit Expressions:* The implicit expressions of like or dislike are involved in the observed disagreement, however the explicit ones – which directly refer to the process of like-dislike expressed by verbs as "like", "love", "enjoy", "hate" – seem more consensual. While the implicit expressions are numerous in the evaluation corpus, the explicit ones are quite rare. An overview of the reference shows that only 21 pairs (over the 496) comprise an explicit verb of likes or dislikes with the user as source. However, these expressions appear as less ambiguous: for 18 of these pairs, the system and the reference agree on the presence of a like-dislike expression. Moreover, this agreement also occurs between the human annotators: over the 25 pairs containing a verb referring to a like-dislike process, 21 show a consensus regarding the presence of a linguistic clue of the user's likes and dislikes.

The agreement on the explicit expressions of likes and dislikes and the observed disagreement on some implicit ones questions the pertinence of considering implicit expressions during the detection. From our point of view, it seems necessary that the detection does not only rely on explicit expressions, which are quite rarely uttered. The detection of the user's implicit like-dislike appears as a useful step for collecting a maximum amount of information on the user's liking. However, in order to distinguish between the explicit expressions and the implicit ones, it will be interesting to notify this explicit or implicit nature in the output provided to the ECA system. Moreover, in order to improve the analysis, the linguistic rules should be able to distinguish between relevant and non relevant implicit expressions. Indeed, if some of them appears as important clues of the user's liking – like those described above –, other seems less suitable. In a pair as, Agent: "oh" – User: "yeah, so its all good", according to the lack of information provided by the context, it should be preferable to prevent the detection of the expression "it's all good".

### *B. Silence and Errors Due to the Conversational and Spontaneous Speech*

If part of the disagreement between the system and the reference relies on the subjective nature of the sentiment-related phenomena, some of its silences and errors can be also due to the processing of spontaneous and conversational speech.

The first difficulty concerns the presence of disfluencies in the sentences. The Semaine corpus contains a great number of utterances where the syntactic structure is disrupted, which can hinder detection. For example, in a pair as, Agent:'Oh!" – User: "are just very good really good film and read a book", if the system is able to detect, at the chunking level, that "very good" and "really good film" are both expressing a positive appreciation, the analysis fails at the sentence level: due to the absence of a subject, the sentence does not match with any relevant attitudinal patterns. Moreover, in this case, no semantic information can be found in the agent's sentence for helping the analysis of the user's like. To handle this difficulty, it would be interesting to integrate a system able to automatically label disfluencies, such as the one presented in [40]. The disfluent structure of the sentence could be integrated to our syntactic and semantic rules.

As a first step for dealing with the conversational context, our system focuses on adjacency pairs. Nevertheless, some agent's utterances are not taken into account in this current version of the system. Regarding the interrogative sentence, the system only deals with yes-no questions. This is the reason for some silences made by the system. In a pair as, Agent: "what kind of things frustrate you during your day" – User: "things like you know not getting things published you know", while the human annotators show a consensus regarding the presence of a user's dislike, the system does not analyse the expression. The attitudinal verb and the wh-question is correctly detected, however, the system is not yet able to analyse the answer. If the yes-no questions uttered by the agent enable a detection of several user's like-dislike expressions, the future version of the system could deal with the wh-questions for improving the results.

Besides, some silences and errors are due to the lack of information provided by some adjacency pairs. A majority of them enable a correct analysis of the user's likes and dislikes, however, few of them make it difficult: the expression is either not detected or misunderstood. For example, in the following pair, Agent: "good. ah good" – User: "my favourite emotion", the system identifies the nominal chunk "my favorite emotion" as a clue of a user's like. However, as no nominal chunk can be identified as a candidate target, either in the user's sentence or in the agent's one, the system does not continue the analysis and the detection fails. An interesting solution will be to design dialogue-based rules dealing with the history of theb conversation. In this way, when the adjacency pair does not provide enough information, the detection could also use the anterior speech turns.

## V. Conclusion and Future Work

In this paper, we introduce a SA method embedded in a ECA system and which deals with the human-agent interaction issues. In this way, we delimit the relevant sentiment-related phenomenon according the agent's communicational goals and make a connection between the liking dimension of Heider [17] and the model of attitude provided by Martin and White [26]. Then, we introduce the ability of our system to deal with the conversational context by modelling agent's utterances for the detection of user's expressions of likes and dislikes. The system is next evaluated by a protocol integrating conversational issues and which relies on a non-expert and system-oriented annotation task. According to the complexity of the targeted phenomenon, the evaluation provides optimistic results for this pioneering version. The in-depth analysis of the agreement and disagreement between the system and the reference provides some tracks for the improvement of the system and a better adaptation to the agent communicational goals: a distinction of implicit and explicit expressions of likes and dislikes, an integration of a larger conversational context and a management of the speech disfluencies.

## Acknowledgment

The authors would like to thank Florian Pecune for Heider's balance theory. This work has been supported by the european project ARIA-VALUSPA, and performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

## References

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

[2] L. Zhang, J. Barnden, R.J. Hendley, M.G. Lee, A.M. Wallington, and Zhigang Wen, "Affect detection and metaphor in e-drama," *International Journal of Continuing Engineering Education and Life-Long Learning*, vol. 18, no. 2, pp. 234, 2008.

[3] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "User study on affectim, an avatar-based instant messaging system employing rule-based affect sensing from text," *International Journal of Human-Computer Studies*, vol. 68, no. 7, pp. 432–450, 2010.

[4] C. Smith, N. Crook, S. Dobnik, and D. Charlton, "Interaction strategies for an affective conversational agent," in *Presence: Teleoperators and Virtual Environments*. 2011, vol. 20, pp. 395–411, MIT Press.

[5] Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity," in *Proceedings of ACL*, 2004, pp. 271–278.

[6] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa, "Just how mad are you? finding strong and weak opinion clauses," in *Proceedings of the 19th National Conference on Artifical Intelligence*. 2004, AAAI'04, pp. 761–767, AAAI Press.

[7] Janyce Wiebe and Ellen Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, Berlin, Heidelberg, 2005, CICLing'05, pp. 486–497, Springer-Verlag.

[8] E. Breck, Y. Choi, and C. Cardie, "Identifying expressions of opinion in context," in *International Joint Conference On Artifical Intelligence*, Sangal S., Mehta H., and Bagga R. K., Eds., San Francisco, CA, 2007, pp. 2683–2688, Morgan KoffMann Publishers.

[9] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Recognition of affect, judgment, and appreciation in text," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Stroudsburg, PA, USA, 2010, COLING '10, pp. 806–814, Association for Computational Linguistics.

[10] K. Moilanen and S. Pulman, "Sentiment composition," in *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, September 27-29 2007, pp. 378–382.

[11] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Citeseer, 2013, vol. 1631, p. 1642.

[12] B. Yang and C. Cardie, "Joint inference for fine-grained opinion extraction," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria, August 2013, pp. 1640–1649, Association for Computational Linguistics.

[13] C. Clavel and Z. Callejas, "Sentiment analysis: from opinion mining to human-agent interaction," *Affective Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, to appear 2015.

[14] C. Langlet and C. Clavel, "Improving social relationships in face-to-face human-agent interactions: when the agent wants to know user's likes and dislikes," in *The Association for Computer Linguistics*, Beijin, China, July 2015, to appear.

[15] João Dias and Ana Paiva, "I want to be your friend: Establishing relations with emotionally intelligent agents," in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multi-Agent Systems, 2013, pp. 777–784.

[16] Magalie Ochs, Nicolas Sabouret, and Vincent Corruble, "Simulation of the dynamics of nonplayer characters' emotions and social relations in games," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 1, no. 4, pp. 281–297, 2009.

[17] F. Heider, *The psychology of interpersonal relations*, Lawrence Erlbaum associates Inc, 1958.

[18] R.B. Zajonc, "The psychology of interpersonal relations," *Public Opinion Quarterly*, vol. 24(2), pp. 280–296, 1960.

[19] C. Osgood, W. H. Mai, and M. S. Miron, *Cross-cultural Universals of Affective Meaning*, University of Illinois Press, Urbana, 1975.

[20] R. Quirk, *A Comprehensive grammar of the English language*, General Grammar Series. Longman, 1985.

[21] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, "Textual affect sensing for sociable and expressive online communication," in *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, Berlin, Heidelberg, 2007, ACII '07, pp. 218–229, Springer-Verlag.

[22] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, "Emoheart: Conveying emotions in second life based on affect sensing from text," *Adv. in Hum.-Comp. Int.*, vol. 2010, Jan. 2010.

[23] C. E. Izard, *Human Emotions*, Plenum Press, New York, USA, 1977.

[24] M. Ishizuka, "Textual affect sensing and affect communication," in *IEEE Transaction 11th International Conference on Cognitive Informatics and Cognitive Computing*, 2012, pp. 2–3.

[25] A. Ortony, G.L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge, University Press, 1990.

[26] J. R. Martin and P. R. White, *The Language of Evaluation. Appraisal in English*, Macmillan Basingstoke, London and New York, 2005.

[27] K. Bloom, N. Garg, and S. Argamon, "Extracting appraisal expressions," *HLT-NAACL*, pp. 165–192, April 2007.

[28] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal taxonomies for sentiment analysis," *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, April 2005.

[29] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2005, HLT '05, pp. 355–362, Association for Computational Linguistics.

[30] C. Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2009, EMNLP '09, Association for Computational Linguistics.

[31] R. Snow, B. O'Connor, D. Jurafsky, and Y. Andrew, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2008, EMNLP '08, Association for Computational Linguistics.

[32] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, June 2011.

[33] C. Langlet and C. Clavel, "Modelling user's attitudinal reactions to the agent utterances: focus on the verbal content," in *5th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals (ES3 2014)*, Reykjavik, Iceland, May 2014.

[34] H. Sacks, E.A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 9-10, pp. 696–735, November 1974.

[35] Elisabetta Bevacqua, Ken Prepin, Radoslaw Niewiadomski, Etienne de Sevin, and Catherine Pelachaud, "Greta: Towards an interactive conversational virtual companion," *Artificial Companions in Society: perspectives on the Present and Future*, pp. 143–156, 2010.

[36] R. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 1083–1086.

[37] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan-March 2011.

[38] J.L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[39] L.J Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[40] C. Dutrey, C. Clavel, S. Rosset, I. Vasilescu, and M. Adda-Decker, "A crf-based approach to automatic disfluency detection in a french call-centre corpus," in *Interspeech*, 2014, p. to appear.