Improving social relationships in face-to-face human-agent interactions: when the agent wants to know user's likes and dislikes

Caroline Langlet
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
caroline.langlet@
telecom-paristech.fr

Chloé Clavel
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
chloe.clavel
@telecom-paristech.fr

Abstract

This paper tackles the issue of the detection of user's verbal expressions of likes and dislikes in a human-agent interaction. We present a system grounded on the theoretical framework provided by (Martin and White, 2005) that integrates the interaction context by jointly processing agent's and user's utterances. It is designed as a rule-based and bottom-up process based on a symbolic representation of the structure of the sentence. This article also describes the annotation campaign – carried out through Amazon Mechanical Turk – for the creation of the evaluation dataset. Finally, we present all measures for rating agreement between our system and the human reference and obtain agreement scores that are equal or higher than substantial agreements.

1 Introduction

In the research field of the embodied conversational agents (ECA), detecting sentiment-related phenomena [1] appears as a key task to improve human-agent interactions and to build long-term social relationships (Pecune et al., 2013). Several models and applications have been proposed which mostly take into account non-verbal cues (acoustic features, facial or bodily expressions) to determine the user's emotions (Schuller et al., 2011). The verbal content is more and more integrated but still partially exploited in human-agent interactions. The very infrequent works, integrating the detection of user's sentiments in ECAs based on linguistic cues, concern avatars and visualisation issues rather than face-to-face interaction, (Zhang et al., 2008; Neviarouskaya et al., 2010b). We identify so far two studies that integrate a sentiment detection module for human-agent interaction (Smith et al., 2011; Yildirim et al., 2011).

However, the research field of sentiment analysis and opinion mining provides a set of interesting works dealing with the subjective information conveyed by the verbal content. Three types of approaches are considered: machine-learning, rule-based approaches and hybrid approaches that are a combination of the first two types. Machine learning methods have proven their worth for the positive and negative classification of sentences or texts (Pang and Lee, 2008). Rule-based approaches are grounded on syntactic and semantic analyses of the sentence and provide deeper analyses of sentiment-related phenomena. For example, (Neviarouskaya et al., 2010a) and (Moilanen and Pulman, 2007) provide linguistic rules dealing with the principle of compositionality in order to improve the detection of opinion targets and the resolution of polarity. Similarly, (Shaikh et al., 2009) provide a linguistic adaptation of the OCC model (Ortony, Clore and Collins (Ortony et al., 1990) based on logic and semantic rules. Hybrid approaches also begin to be used for more fine-grained opinion and sentiment analysis (Yang and Cardie, 2013)

Sentiment/opinion detection methods used in human-agent interaction are rare and, when they are employed, they are not different from the ones used in opinion mining: they are consequently not designed for socio-affective interactions. Indeed, the development of a module for the detection of sentiment-related phenomena in face-to-face human-agent interactions requires to tackle

[1] The term *sentiment-related phenomena* is used in (Clavel et al., 2013) to regroup all the phenomena related to sentiment in the literature, from opinion to affect and emotion.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1064–1073,
Beijing, China, July 26-31, 2015.

various scientific issues: the delimitation of the relevant sentiment-phenomenon to detect, the integration of the multi-modal context and the management of the spontaneous and conversational speech.

The present paper tackles two of the issues: the integration of the conversational context and the delimitation of the relevant phenomenon. Regarding the first issue, we propose a system relying on a rule-based method that allows us to model the agent's utterances in order to help the detection of user's sentiment-related phenomena.

Then, we delimit and specify the linguistic phenomenon to detect by focusing on one specific aspect required by ECAs for modelling social relationships: the user's likings that are given by the expressions of user's likes and dislikes in the verbal content.

This paper is organised as follows: first, we present the theoretical model which our system is grounded on (Section 2). Then, we provide a description of the system: each stage of the bottom-up process is described, including the linguistic rules and the patterns used by the system. In Section 4, we introduce the annotation campaign we launched on Amazon Mechanical Turk (AMT) in order to create a data-set for the evaluation of our system. Finally, we present and discuss the results of the system evaluation (Section 5).

2 Theoretical background

The *liking* is one of the key dimensions used for the modelling of social relationships (Pecune et al., 2013). The definition of this concept is grounded on the Heider's *Balance Theory* (Heider, 1958) and is defined as: "the way relations among persons involving some impersonal entity are cognitively experienced by the individual" (Zajonc, 1960). Heider's theory is integrated in social agent computational models by defining scenarios where the agent and the user's likings toward each other are determined by their liking toward other entities (things, process or events). In such scenarios, the analysis of user's verbal content has a key role as a major source of information for determining of the user's likes and dislikes. Therefore, a linguistic description of this phenomenon is required to design a detection system.

In the research field of *Opinion Mining* and *Sentiment Analysis*, the majority of opinion/sentiment detection systems focus on the positive/negative distinction or on the classification of a restricted number of emotion categories. Other in-depth approaches, as (Wiebe et al., 2005; Breck et al., 2007), refer to the Private State Theory, which defines mental states as involving opinions, beliefs, judgements, appraisals and affects. Beside those models, the model proposed by (Martin and White, 2005) is increasingly used in several works (Neviarouskaya et al., 2010a; Bloom et al., 2007; Whitelaw et al., 2005). This model provides a linguistic description and a focus on the verbal expressions of sentiment and opinion and proposes a complex framework, for describing how *attitudes* are expressed in English. It distinguishes affects – which are concerned with emotional reactions – from judgements and appreciations – which relate to evaluations toward people's behaviours and semiotic or natural phenomena. Finally, it models attitudinal expressions as relying on three elements: a source, the person evaluating or experiencing, a target, the entity which is evaluated or which triggers an affect and a linguistic clue expressing the evaluation.

In this model, likes and dislikes can be considered as a subcategory of the *Attitudes*. This subcategory overlaps the three categories (affect, judgment, appreciation) defined by (Martin and White, 2005). For example, the sentence *"This painting makes me sad"* is considered as an affect, while the sentence *"This painting is a master-work"* is considered as an appreciation. But, in both cases, we can consider them as a user's like. However, among the expressions of attitudes where the source is the user, some of them do not refer to a like or dislike. For example, *"I'm very happy"* refers to an affect and does not give any clue regarding a possible like or dislike. Thus, a selection of relevant attitudes have to be done. The rules used for this selection are presented in the next section.

3 A rule-based and symbolic method

On the basis of the Martin and White's model described in the previous section, we design a system able to detect expressions of attitudes corresponding to the user's likes and dislikes. It is grounded on linguistic rules modelling the syntactic and semantic structure of the sentences.

3.1 Integrating the interaction context

The system presented in Figure 1 successively processes each adjacency pair (AP) of the dialogue

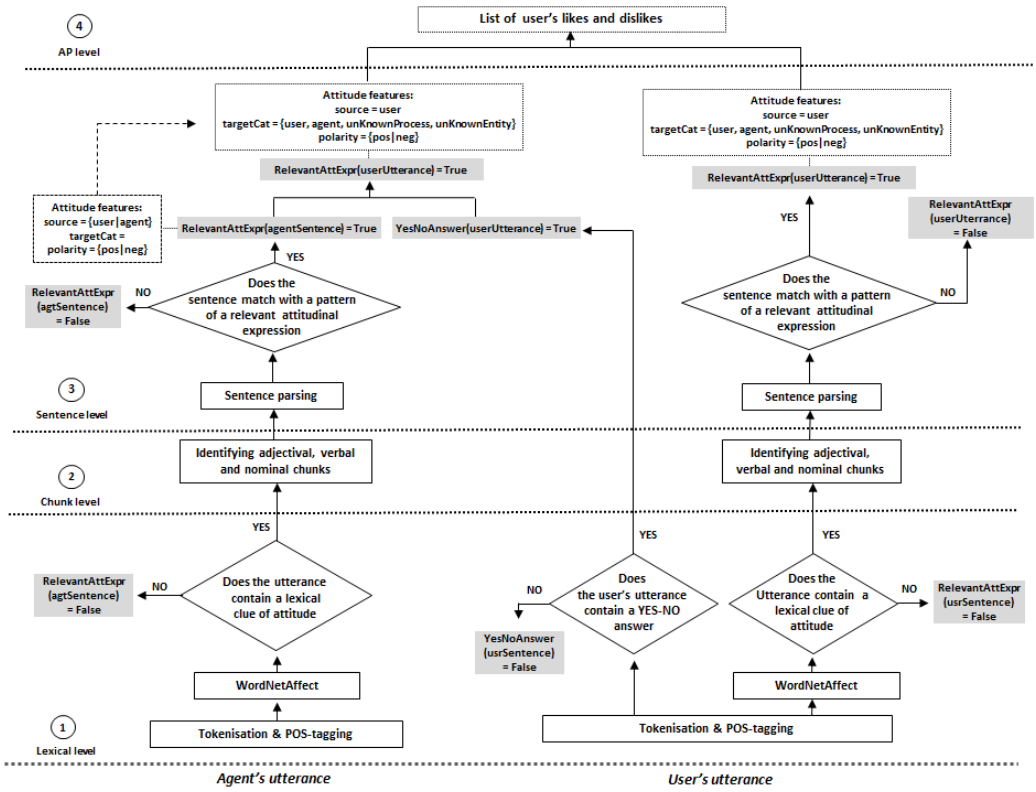


Figure 1: Process overview

(Sacks et al., 1974), i.e. each user's speech turn and the agent's one immediately preceding it. We aim to detect two kinds of user's attitudinal expressions that can occur during the interaction: the first ones which are spontaneous and do not depend on the agent's sentence (Agent: *What did you do today?* User: *I saw a great movie*); and the second ones which are triggered by the agent's sentence (Agent: *Do you like outdoors activities?* User: *Yeah very much*).

In the last case, the detection of the attitude expressed in the agent's sentence appears as a necessary step for the detection of the user's ones. This detection has to be done in an automatic way as, in the agent platform we use (the Greta platform, (Bevacqua et al., 2010), the agent's speech turns are not automatically generated but scripted. Thus, we cannot obtain the linguistic and semantic information about attitude by using the generation data. Furthermore, in order to make the dialogue setting as light as possible, it is not possible to script such values for each agent's sentence.

3.2 A bottom-Up process

The relevant expressions of attitudes are detected by using a bottom-up and rule-based process, which launches successively the different levels of analysis: lexical level, chunk level, sentence level. These three stages comprise formal grammars, which are implemented within the Unitex plateform (Paumier, 2015). During these various stages, values are assigned to the three boolean variables which are finally used to decide whether the user is expressing a like or a dislike: $RelevantAttExpr(agtSentence)$, $RelevantAttExpr(usrSentence)$, $YesNoAnswer(usrSentence)$.

3.2.1 Lexical level

After a tokenisation and a POS-tagging, the system checks whether the sentence (the user or the agent's one) contains lexical clues of attitudinal expressions. Three parts of speech are taken into account: the nouns, the adjectives and the verbs. We use a re-adaptation of the Wordnet-affect lexicon (Valitutti, 2004). In order to adapt this lexicon to our goal, a selection of relevant lexical entries has to be done. Among all the synsets, we select those which can be linked to like and dislike and that belong to the following main categories: *positive-emotion*, *negative-emotion*, *neutral-emotion*. As the lexi-

	Syntactic patterns	Attitude and Polarity rules
		For all chunk: If the chunk conveys a word labelled as a lexical clue of attitude, it is considered as an attitudinal chunk (ex: "a very good movie") `ChkNoun(att:true); ChkAdj(att:true); ChkVb(att:true)`
Adjectival chunk	`ChkAdj → Adv?, Adj`	`ChkAdj(att:true, pol:pol`$_{Adj}$`)` As no modifier can shift the polarity of the adjective in the adjectival chunk, the polarity of the adjectival chunk is always equal to the adjective one (.

Nominal chunk	`ChkNoun → Det?, ChkAdj?, Noun`	`If Noun(Pol`$_{Noun}$`) AND ChkAdj(att:true, pol:inv(Pol`$_{Noun}$`)):`
		(i) if the noun and the adjective have an attitudinal value, and if the their polarity differ, the polarity assigned to the nominal chunk is the same as the adjective (ex: ``performance'', positive, ``bad performance'', negative);
		`If Det == neg :`
		`ChkNoun(att:true, pol:inv(Pol`$_{Noun}$`))`
		(ii) if the determiner has a negative value -- *no*, *any*, for example -- the polarity assigned to the nominal chunk is the opposite of the nominal polarity, otherwise the polarity is the same as the noun (ex: ``any good idea'').

Verbal chunk	`ChkVb → Aux?, Vb.`	`If Neg`$_{Verb}$ `== True OR Neg`$_{Aux}$ `== True :`
		`ChkVerb(att:true, pol:Pol`$_{Noun}$`)`
		if the verb or one of its auxiliaries convey a negation, the polarity assigned to the verbal chunk is the opposite of the verbal polarity (ex: ``don't like''), otherwise the polarity is the same as the verb (``like'').

Figure 2: Patterns and polarity rules used for the chunk level

con is applied by the Unitex plateform, we turn the Wordnet-Affect lexicon into a unitex dictionary format. Finally, this transformation provides three dictionaries: one for the nouns, one for the adjectives and one for the verbs. If the lexical processing has found one or several lexical clues of attitude, the system continues the analysis and get to the next stage, else $RelevantAttExpr(X) = False$ and the system quits the analysis of the sentence. Regarding the user's sentence, the system also checks if one or several tokens of sentence match with a yes or a no word, by using a short lexicon manually built which comprises less than ten words for each sentence type. If the test succeeds $YesNoAnswer(usrSentence) = True$, else $YesNoAnswer(usrSentence) = False$.

3.2.2 Chunk level

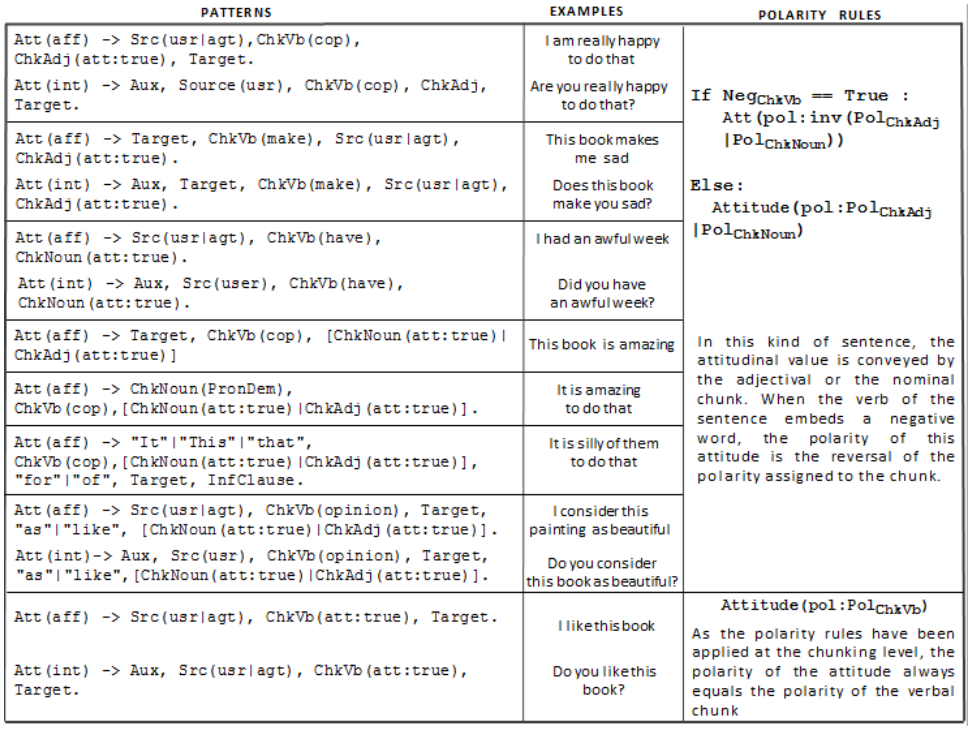
At this level, we design formal grammar – implemented as finite state automatons within the Unitex plateform. Three main chunks are defined: the verbal, the adjectival and the nominal chunks. All these chunks can imply a lexical unit of attitude. In such case, a polarity value is assigned to the entire chunk by applying rules which consider valence shifters and polarity conflict (see Figure 2).

3.2.3 Sentence level

Attitudinal value The system parse of each sentence for checking if the sentence matches with an attitudinal expression, according to its syntactic structure. This parsing phase is grounded on a set of patterns (see Figure 3). Among the attitudinal patterns provided in the literature (Neviarouskaya et al., 2010a; **?**), we selected those expressing a like or dislike according to a previous corpus-based study (Langlet and Clavel, 2014) (development corpus presented in Section 4.1). Depending on the speaker of the processed sentence – the agent or the user – sentence structures can be interrogative or affirmative surface structures. In the agent's sentence, the system looks for both affirmative and interrogative forms, while in the user's sentences, it only takes into account affirmative structures.

Type of the source Simultaneously, the system checks the source of the attitude. The type of a relevant source varies depending on the sentence processed: in the **agent's sentence**, the system aims to detect attitudes able to be validated or invalidated by the user and whose source is either the agent – lexically represented by a first person pronoun ($Src(agt) \rightarrow$ "I"|"me") – or the user – lexically represented by a second person pronoun ($Src(usr) \rightarrow$ "you'); in the **user's sentence**, the system aims to detect only the attitudes whose source is the user – represented by a first person pronoun ($Src(usr) \rightarrow$ "I"|"me").

Target and polarity At this stage, the system is also able to define the polarity of the expression



PATTERNS	EXAMPLES	POLARITY RULES
Att(aff) -> Src(usr|agt),ChkVb(cop), ChkAdj(att:true), Target. Att(int) -> Aux, Source(usr), ChkVb(cop), ChkAdj, Target.	I am really happy to do that Are you really happy to do that?	If Neg_{ChkVb} == True : Att(pol:inv(Pol_{ChkAdj}|$Pol_{ChkNoun}$)) Else: Attitude(pol:Pol_{ChkAdj}|$Pol_{ChkNoun}$)
Att(aff) -> Target, ChkVb(make), Src(usr|agt), ChkAdj(att:true). Att(int) -> Aux, Target, ChkVb(make), Src(usr|agt), ChkAdj(att:true).	This book makes me sad Does this book make you sad?	
Att(aff) -> Src(usr|agt), ChkVb(have), ChkNoun(att:true). Att(int) -> Aux, Src(user), ChkVb(have), ChkNoun(att:true).	I had an awful week Did you have an awful week?	
Att(aff) -> Target, ChkVb(cop), [ChkNoun(att:true)|ChkAdj(att:true)]	This book is amazing	In this kind of sentence, the attitudinal value is conveyed by the adjectival or the nominal chunk. When the verb of the sentence embeds a negative word, the polarity of this attitude is the reversal of the polarity assigned to the chunk.
Att(aff) -> ChkNoun(PronDem), ChkVb(cop),[ChkNoun(att:true)|ChkAdj(att:true)].	It is amazing to do that	
Att(aff) -> "It"|"This"|"that", ChkVb(cop),[ChkNoun(att:true)|ChkAdj(att:true)], "for"|"of", Target, InfClause.	It is silly of them to do that	
Att(aff) -> Src(usr|agt), ChkVb(opinion), Target, "as"|"like", [ChkNoun(att:true)|ChkAdj(att:true)]. Att(int)-> Aux, Src(usr), ChkVb(opinion), Target, "as"|"like",[ChkNoun(att:true)|ChkAdj(att:true)].	I consider this painting as beautiful Do you consider this book as beautiful?	
Att(aff) -> Src(usr|agt), ChkVb(att:true), Target. Att(int) -> Aux, Src(usr|agt), ChkVb(att:true), Target.	I like this book Do you like this book?	Attitude(pol:Pol_{ChkVb}) As the polarity rules have been applied at the chunking level, the polarity of the attitude always equals the polarity of the verbal chunk

Figure 3: Patterns and polarity rules used for the sentence level: the second column presents examples of sentences matching with the patterns detailed in the first column. The rules introduced in the third column are applied according to the sentence pattern detected.

by detecting the valence shifters which can modify the polarity of the attitudinal chunk and by applying the appropriate polarity rules described in Figure 3. Regarding the target, the system is only able to assign to the target one of four generic classes. The first two classes concern the two members of the conversation – *agent* and *user*. The third class, called *other*, deals with all entities and processes which are neither the agent or the user. The last one – *unknown* – concerns all the target referring by a pronoun, and whose class – even generic – cannot be known. In a future work, the *other* category could be detailed by using an ontological resource, and *unknown* category by referring to an anaphora resolution.

3.2.4 User's utterance level within the AP

Generating attitude feature set Once the sentence level is done, the $True$ value is assigned to the $relevantAttExpr(usrSentence)$ variable in two steps.

Firstly, the syntactic structure of the user's sentence matches with one of the attitudinal patterns (Figure 3) whose source is the user ($Src(user)$). The feature set of the attitudinal expression is generated according to the information found at the parsing stage: $source \in \{user, agent\}$, $polarity \in \{neg, pos\}$, $targetType \in \{user, agent, other, unknown\}$.

Secondly, if the agent's sentence matches with one of the attitudinal patterns whose source is either the user or the agent ($Src(user|agent)$), then $relevantAttExpr(agtSentence) = True$. In this second case, if $YesNoAnswer(usrSentence) == True$, the user validates or invalidates the attitude. Thus, the system defines $relevantAttExpr(usrSentence) == True$, even if any sentence matching with a relevant pattern has been found in the user's sentence. The feature set associated to the user's attitude is built according to those assigned to the attitudinal expression found is the agent's sentence. Since the user assumes or rejects the attitude expressed by the agent, the system considers that he/she utters an attitudinal expression that he/she is the source. Regarding the polarity, if the user validates the statement expressed by the agent, the polarity of his/her attitude is the same as the agent's one. Otherwise, if the user expresses a no answer, the polarity is the opposite of the agent's one.

Converting attitude into like-dislike The patterns used for the parsing phase refer to attitudes that are good candidates for expressions of like or dislike. When the $relevantAttExpr(usrSentence) == True$, the system converts the attitude into a like or a dislike on the basis of the feature set associated to the expression of attitude: an attitude with a positive polarity ($attitude(pol : pos)$) is considered as a $like$, and an attitude with a negative polarity ($attitude(pol : neg)$) is considered as a $dislike$. The target is the same as the attitudinal expression.

4 Corpus for evaluating the system

4.1 Semaine corpus

In order to evaluate our system, an annotated data set of sentences extracted from the Semaine corpus (McKeown et al., 2011) has been created. This corpus comprises 65 manually-transcribed sessions where a human user interacts with a human operator playing the role of the virtual agent. These interactions are based on a scenario involving four agent characters: Poppy, happy and outgoing, Prudence, sensible and level-headed, Spike, angry and confrontational and Obadiah, depressive and gloomy. Agent's sentences are constrained by a script (however, some deviations to the script occur in the database) aimed at putting the user in the same state as the one of the played character. 30 sessions of the corpus have been used for the development set. The rest of the data has been considered to build the evaluation corpus following the protocol described in the next paragraph.

4.2 Annotation protocol on AMT

We use AMT platform to carry out the annotation campaign. It allows us to easily recruit a large number of English native speakers. Recent works have shown the reliability of the annotations provided by this platform. For various tasks of language annotation – evaluation of machine translation (Callison-Burch, 2009), affect recognition (Snow et al., 2008), or dictionary validation (Taboada et al., 2011) – they observe a high agreement of non-expert raters with the gold standards.

For our annotation protocol, the recruited annotators are put in the same conditions as the system: each annotator has to label the user's likes and dislikes by only considering the AP (without the whole interaction) and the verbal content (without the audio and video). Among the pairs having less than thirty words in the evaluation corpus, we randomly selected 600 APs – made of an agent's speech turn and a user's one (see Section 3.1). This length of the sentence has been restricted to avoid annotation difficulties.

The dataset is divided in 60 subsets of 10 APs. In order to secure the annotation and to prevent the annotators from doing the annotation task two times, we use TurkGate tool (Goldin and Darlow, 2013). 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. 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).

Questionnaire As the annotation is done by non-expert annotators, we design a simplified and intuitive annotation process: for each pair, the annotators have to answer to a set of questions featured in Figure 4. The goal of the questionnaire is to determine whether the annotator is able to deduce a user's like or dislike from the APs. 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 to the place of the hearer. Each question of the questionnaire focuses on one of the outputs of the detection system:

- The **first question** examines the presence of an expression of like or dislike and provides a *yes*/*no* answer.
- the **second question** deals with the multiple occurrences of like/dislike expressions in the same speech turn. We limited the answer to "4" (maximum number of like/dislike expressions observed in the dataset). If the annotator detects more than one expression of like/dislike, the questions 3 to 4 are asked for each expression of like/dislike.
- the **third question** deals with the type of the target. As answers, only the four types –

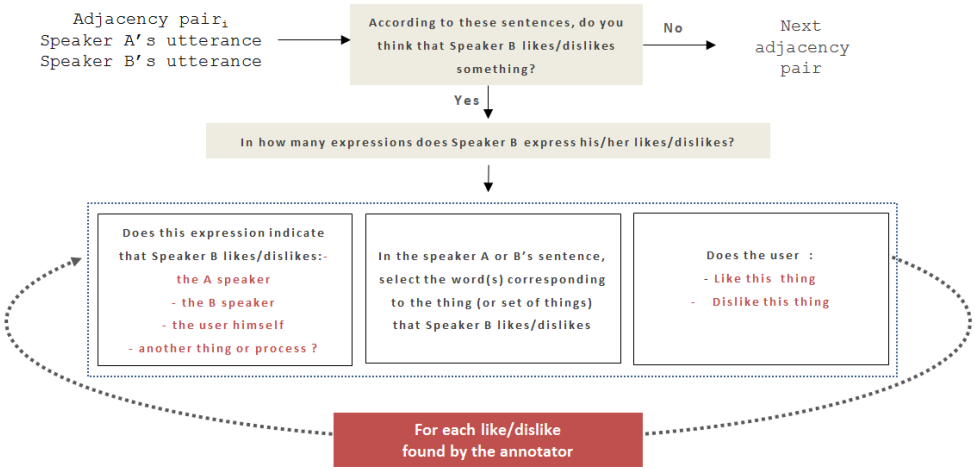


Figure 4: Annotation process on AMT

those the system is able to detect – are proposed.

- The **fourth question** concerns the polarity of the expression: positive (like) or negative (dislike).

4.3 Inter-annotator agreement and consistency

We measure the inter-annotators agreement or consistency at each stage of the questionnaire. All the measures presented in the section have been applied for each subset of the corpus (60 subsets of 10 APs, 4 annotators for each subset).

	Fleiss' Kappa	**Cronbach's alpha**
Max	0.79	0.90
Median	0.32	0.72
Average	0.25	0.59

Table 1: Fleiss' kappa scores and Cronbach's alpha coefficients obtained in on the 60 subsets

Regarding the answer to the first question of the questionnaire, we measure how the annotators are agreeing on the presence of at least one user expression of like or dislike by using the Fleiss' Kappa (Fleiss, 1971) (see Table 1). Second, we measure the consistency on the annotation of the number of user's expressions to each pair by using the Cronbach's alpha coefficient (Cronbach, 1951). As, for labeling the number of likes-dislikes expressed in each pair, the crowd-workers have to select a value on a scale (from 1 to 4), it appears as suitable to measure the relative similarity between ratings rather than the agreement about an exact value. The Cronbach's alpha is designed for evaluating the internal consistency of a scale annotation. In this way, it measures the degree to which different raters or observers make consistent estimates of the same phenomenon.

The obtained scores are encouraging. Regarding the agreement on the presence of an expression of like or dislike, even if the median score is comprised between 0.30 and 0.40, the maximal value equals to 0.79. Moreover, 40% of the subsets has a kappa score comprised between 0.40 and 0.60. The consistency score is also significant: 51% of the annotated sub-corpus has a score equal or higher than 0.7, which is considered as an acceptable level of agreement (George and Mallery, 2010).

For the polarity and the target type, we select the pairs where at least two annotators agree on the presence of an expression of like or dislike, and we consider only the annotations provided by these annotators. After this selection, we obtain a sub-set of ratings with a unfixed set of annotators. As the Fleiss' Kappa must be applied on data with an invariable and fixed set of raters, we consider the percent agreement (Gwet, 2010) as more appropriate. Even though, it seems sometimes difficult for the annotators to agree on the presence of a user's expression of like or dislike, their agreement on the polarity of such expressions appears as more significant: 41% of the sub-corpus has a percentage of agreement between 50% and 75% and 52% of the sub-corpus has a percentage of agreement upper than 75%. The agreement is

also significant regarding the target: 61% of the sub-corpus has a percentage of agreement upper than 50%. All these results are quite positive for a system-oriented annotation of a such subjective phenomenon.

5 Evaluation of the system

5.1 Protocol

From the 600 pairs of the previously annotated corpus, we keep 503 pairs for the evaluation of the system by removing the pairs where a consensus can not be found between the 4 annotators – that is that we keep as a reference the majority vote corresponding to the data where at least three annotators agree. We use three different measures to evaluate the system performance relying on the agreement measures presented in Section 4.3: the detection of the presence of a user's expression of like or dislike is evaluated by the Fleiss' kappa between the system output and the reference; the consistency on the number of detected expressions is evaluated by the Cronbach's alpha coefficient; the agreement on the polarity is measured by using the Fleiss' kappa; and the agreement on the target type with the percentage of agreement.

5.2 Results

Table 2 presents the results obtained for each detection task (presence of a like/dislike expression, detection of the correct number of expressions contained in an sentence, and correct classification between like and dislike). The agreement between

No Expr-Expr	$k = 0.61$
Nb of expressions rated	$\alpha = 0.67$
Polarity	$k = 0.84$
Target type	$p = 53\%$

Table 2: Agreement scores between the system output and the reference

the system output and the reference is substantial for the detection of the presence of a user expression ($k = 0.61$) and the number of user expressions is also correctly detected by the system (acceptable α largely higher than 0.6). However, the major part of the corpus contains no more than 1 like/dislike expression (98% of the pairs are annotated by the reference and the system as containing 0 or 1 like/dislike expression). 4% of the pairs (25 pairs) is annotated by the system as containing 1 expression, while the referred annotation does not indicate the presence of any like/dislike expression. For 8% of the pairs (43 pairs), it is the opposite phenomenon (1 expression annotated by the reference but not by the system). The Fleiss' kappa score obtained for the polarity is really encouraging since it equals 0.844. Regarding the target type, we obtain a percentage of agreement at 53%. The disagreement frequently concerns a confusion between the *unknown* and *other* categories.

5.3 Discussion

We have carried out an in-depth analysis of the disagreement between the system outputs and the human annotations in order to identify tracks for the improvement of the system. We identified two main types of difficulties.

The first difficulty concerns the processing of spontaneous speech. The Semaine corpus contains a great number of disfluent utterances that disrupt the syntactical structure of the speech turn and thus hinder both the annotation process and the detection system. In the following pair, Agent: *''Oh!"* – User: *''are just very good really good film and read a book"*, the grammatical structure of the user sentence is fuzzy (absence of the subject, presence of repairs) which makes the parsing of the sentence and thus the detection of attitudinal patterns difficult. However, the annotators have here correctly identified the presence of a like and the type of the target ("the film" in the *Others* category), which is not the case for all the annotations of the disfluent utterances. To handle this difficulty, it would be interesting to integrate a system able to automatically label disfluencies, such as the one presented in (Dutrey et al., 2014). The disfluent structure of the sentence could thus be integrated to our syntactic and semantic rules. However, the automatic detection of disfluencies is still an open challenge, in particular in the case of edit disfluencies where the speaker corrects or alters the utterance or abandons it entirely and starts over (Strassel, 2004).

The second difficulty concerns the lack of context provided by some of the APs. Our system offers a first step in the integration of the interaction context by considering jointly the user's utterance and the previous agent's one that allow us to correctly analyse a large scale of expressions. However, the system and the annotators have to focus on the APs without considering the preceding

speech turns, which can cause disagreements not only between the system outputs and the human annotations, but also between the human annotators. In the following example, Agent: *"good. ah good"* – User: *"my favourite emotion"*, the source (here, the user) can be easily identified but the information contained in the AP is not sufficient to identify the target. An interesting answer to this issue is to take into account the whole conversation preceding a user's utterance as a significant context for the latter. This will imply the design of new complex rules taking into account a larger interaction context.

6 Conclusion and future works

We have introduced a NLP-based system able to detect user's expressions of likes and dislikes in the conversation with an ECA. This system relies on syntactic and semantic rules integrating the interaction context by analysing the content of the agent's utterances to help the analysis of the user's ones. It is designed as a bottom-up and rule-based process. The system has been evaluated by using an evaluation data set created under AMT platform. This first and pioneering version of the system shows encouraging results for the different tasks performed by the system that concern the detection of relevant like/dislike expressions (substantial agreement with a Fleiss kappa at 0.61), the categorization of the expressions between like and dislike (almost perfect agreement with a Fleiss kappa at 0.84) – polarity assignment – and the identification of the target type (53% of agreement between the reference and the system output). Beyond these quite optimistic results, we have provided some tracks for the system improvement that concerns a deeper integration of the interaction context and the processing of spontaneous speech features.

Acknowledgments

The authors would like to thank the GRETA team for its contributions to the Greta and Vib platforms. 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

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

K. Bloom, N. Garg, and S. Argamon. 2007. Extracting appraisal expressions. *HLT-NAACL*, pages 165–192, April.

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

C. Callison-Burch. 2009. 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*, EMNLP '09, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Clavel, C. Pelachaud, and M. Ochs. 2013. User's sentiment analysis in face-to-face human-agent interactions prospects. In *Workshop on Affective Social Signal Computing, Satellite of Interspeech*. Association for Computational Linguistics, August.

L.J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

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

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

D. George and P. Mallery. 2010. *SPSS for Windows Step by Step: A Simple Guide and Reference 18.0 Update*. Prentice Hall Press, Upper Saddle River, NJ, USA, 11th edition.

G. Goldin and A. Darlow. 2013. Turkgate (version 0.4.0) [software]. `http://gideongoldin.github.com/TurkGate/`.

K. L. Gwet. 2010. *Handbook of Inter-Rater Reliability*. 11th edition.

F. Heider. 1958. *The psychology of interpersonal relations*. Lawrence Erlbaum associates Inc.

C. Langlet and C. Clavel. 2014. 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.

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

G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, Jan-March.

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

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

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. 2010b. User study on AffectIM, an avatar-based Instant Messaging system employing rule-based affect sensing from text. *International Journal of Human-Computer Studies*, 68(7):432–450.

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

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January.

S. Paumier. 2015. *Unitex user manual*. Université de Paris-Est Marne-la-Vallée.

F. Pecune, M. Ochs, and C. Pelachaud. 2013. A formal model of social relations for articial companions. In *EUMAS 2013*, December.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(9-10):696–735, November.

B. Schuller, A. Batliner, S. Steidl, and D. Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, November.

M. Shaikh, H. Prendinger, and M. Ishizuka. 2009. A linguistic interpretation of the occ emotion model for affect sensing from text. In *Affective Information Processing*, pages 378–382. Springer London, May.

C. Smith, N. Crook, S. Dobnik, and D. Charlton. 2011. Interaction strategies for an affective conversational agent. In *Presence: Teleoperators and Virtual Environments*, volume 20, pages 395–411. MIT Press.

R. Snow, B. O'Connor, D. Jurafsky, and Y. Andrew. 2008. 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*, EMNLP '08, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Strassel, 2004. *Simple Metadata Annotation Specification*. Linguistic Data Consortium.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), June.

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

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

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotation expressions of opinion and emotions in language. *Language Resources and Evaluation*, pages 165–210, Vol. 39/2-3.

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

S. Yildirim, S. Narayanan, and A. Potamianos. 2011. Detecting emotional state of a child in a conversational computer game. *Computer Speech and Language*, 25(1):29–44.

R.B. Zajonc. 1960. The psychology of interpersonal relations. *Public Opinion Quarterly*, 24(2).

L. Zhang, J. Barnden, R.J. Hendley, M.G. Lee, A.M. Wallington, and Zhigang Wen. 2008. Affect detection and metaphor in e-drama. *International Journal of Continuing Engineering Education and Life-Long Learning*, 18(2):234.