Using Context to Disambiguate Communicative Signals

Mark ter Maat and Dirk Heylen

Human Media Interaction, University of Twente
PO Box 217, 7500 AE Enschede
The Netherlands
{maatm,d.k.j.heylen}@ewi.utwente.nl

Abstract. After perceiving multi-modal behaviour from a user or agent a conversational agent needs to be able to determine what was intended with that behaviour. Contextual variables play an important role in this process. We discuss the concept of context and its role in interpretation, analysing a number of examples. We show how in these cases contextual variables are needed to disambiguate multi-modal behaviours. Finally we present some basic categories in which these contextual variables can be divided.

Keywords: Communicative signals, functions, intentions, context, disambiguation, contextual elements.

1 Disambiguating Signals

In the world of virtual, agents Embodied Conversational Agents (ECA's) take a special place [3]. These agents are capable of detecting and understanding multi-modal behaviour of a user, reason about it, determine what the most appropriate multi-modal response is and act on this.

The SAIBA framework has been proposed as a general reference framework for ECA's in a collaboration of several research groups[1]. In this framework the communicative function, the high-level intention, is separated from the actual multi-modal behaviour (a communicative signal) that is used to express the communicative function. For example, the function 'request turn' is conceptually separated from the act of raising a hand and breathing in to signal that you want to speak. This separation is realized by putting the responsibilities of the functions and signals in different modules. As a result, the modules in the system should be capable of communicating these functions and signals to each other. This communication is performed using specification languages called FML, Function Markup Language, which is used to specify the conversational function, and BML, Behaviour Markup Language, which is the high level specification of the actual multi-modal behaviour. In the current debate about FML an important issue that is being raised is what constitutes a function and how should they be distinguished from contextual elements [8].

[1] http://wiki.mindmakers.org/projects:saiba:main/

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 67–74, 2009.
© Springer-Verlag Berlin Heidelberg 2009

Contextual elements consist of all the elements that surround the communicative behaviour that is being analyzed. This includes, for example, facial expressions that appear at the same time, the fact that the previous sentence was a question, or the fact that the conversation is taking place in a noisy environment.

Ideally an ECA is capable of perceiving and interpreting the multi-modal behaviours of a human user (or another virtual human): the speech, gestures, facial expressions etc. We call the multi-modal behaviours that an ECA can perceive (detected) "signals". The intelligent agent will try to find out what was meant by the behaviours, which means trying to map the signals on the communicative functions that lie behind them.

As is well-known, determining the intended communicative function from detected signals is not a trivial problem as the same signal may occur in different contexts with widely diverging meanings. For example, if you detect that the other person (who is currently speaking) is suddenly gazing at you, then this might be part of a behaviour complex that signals the intention of a turn offer [12,2] or it may indicate a request for feedback [1,8,10]. In many cases it is immediately clear from the context why the behaviour was performed. But what is context? And what kind of elements does it consists of?

In this article we analyse the notion of context, the influence it has on behaviour selection and behaviour interpretation, and the overlap it has with the definition of signals. Finally, a rough categorisation of the elements that should appear in the context is presented. But first some background information about the use of context is provided.

2 Background

The notion of context in interpretation is an important topic in the literature. In this section we briefly discuss three perspectives on context. The first from the linguistic literature ([9] and [7]). The second from the literature on nonverbal communication ([5]) and the last from the ECA community ([11]).

In [9] the author describes several linguistic situations in which the context should be used, and these situations appear on different levels. For example, on the syntactic level, words can have multiple lexical categories. A word like *phone* can be a noun ('Please pick up the phone') or a verb ('I will phone you tomorrow'). The surrounding words (forming the context of the word *phone*) have to be used to determine the exact lexical class of the word.

On a semantic level the same situation occurs. Also the meaning of a single word also depends on the context. And on an even higher level, the grammatical mood of a sentence also depends on the context. An utterance such as 'I want you to come here' (example taken from [9]) is mostly classified as a declarative sentence, but under the right circumstances (say, a mother that, after a few attempts, tries to call her child with this sentence spoken in a very fierce voice) can also be classified as a command (an imperative sentence).

In [7], consider various aspects of what they call the *context of situation*. This is a reference to the environment in which discourse is situated, and to the extra-linguistic factors that have a relation to the text produced itself. This

environment is used to put the text into perspective. For example (taken from [7]) a set of non-cohesive sentences might suddenly become cohesive when it appears in a language textbook as an example set.

According to the authors, the context of situation can be described with three different components: the field, the mode and the tenor. The field is the current topic that is being discussed, the mode is about the communication channel (type of channel, the genre of the interaction), and the tenor is about the participants in the discourse and their relationships to each other. They argue that by describing the context of situation with these three parameters a text can be understood correctly as a coherent passage of discourse.

But not only text needs context. In [5] the authors discuss various aspects concerning the interpretation of non-verbal behaviour. One element of their analysis is the usage of the behaviour, which means looking at the regular and consistent circumstances that surround the non-verbal acts, which basically is the context of the act. These circumstances can be grouped into several categories:

External condition – refers to environmental circumstances, for example the setting (home, work, interview, etc).

Relationship to verbal behaviour – specifies the relationship of the non-verbal with the verbal behaviour, for example if the non-verbal behaviour repeats, augments, accents or contradicts certain words.

Awareness – defines whether the person knows he is performing a particular non-verbal act at the moment he does it.

Intentionality – specifies whether the actor does the act deliberately or not.

External feedback – defines signals that the listener sends back to the speaker to acknowledge the fact that he perceives and evaluates the speaker's actions.

Type of information conveyed – refers to the different types of non-verbal behaviour.

- **Informative** – an act (which may or may not be intentional) that provides the listener with some information. This means that the act at least bears some meaning for the listener.
- **Communicative** – a consciously intended act by the sender to transmit a specific message to the listener
- **Interactive** – an act by one person in an interaction which clearly tries to modify or influence the interactive behaviour of the other person(s).

In [11] the authors analyse the context of multi-modal behaviour of interactive virtual agents. They argue that mapping FML to BML, in other words, choosing the concrete behaviour to perform as an agent when a conversational function is provided, cannot be done without any context. For example, saying hello in the morning ('good morning') is different than in the evening ('good evening'), and this difference can only be made when the local time is known. To add context to the mapping they suggest a new representation, namely CML: Context Markup Language. This language is specifically created to communicate context and consists of three parts: the Dialogue context, the Environmental context and the Cultural context. The dialogue context includes the history of the dialogue, the current topic, the level of tension between characters, etc. The environmental

context contains such things as the time of day, the current setting, etc. The cultural context contains "information on the culturally appropriate way to express certain communicative functions".

3 Towards Contextual Elements

In this section the notion of 'context' is discussed, with the intention of finding out what exactly is context and what it consists of. At the end of this section the elements that make up the context (and are used to disambiguate signals) are categorized into three types. To find these elements and categories, this section starts with the question of how to delimitate what constitutes a signal. What elements are part of a signal and which are part of the context? Next the effect that context has in a conversation is discussed and the problem of ambiguity is addressed.

3.1 Context or Part of the Signal

An important distinction to make when talking about signals and their context is what exactly should be defined as context and what should be considered part of a signal.

What constitutes a particular signal is hard to define as this faces two problems: segmentation and classification. The first is the act of determining what elements are part of a particular signal. As we are trying to associate signals with meanings, the question is "What are the meaning-bearing units that we are talking about?" The input of a conversational agent usually is a constant stream of multi-modal behaviours (position and movement of limbs, spoken words, etc). To identify signals in this stream, it has to be cut at certain boundaries. But segmentation not only consists of determining the boundaries of a signal, but also determining which elements within those boundaries are parts of a signal. Is the movement of the eyebrows a single signal? Or should that signal include the complete facial expression? And should the position or the movement of the head be taken into account as well?

The problem of classification arises when a particular signal has been identified. Just as linguistic phonemes may infinitely differ in their phonetic realisation, the realisation of the non-verbal equivalent of a phoneme will each time be realised in a specific ways along various parameters. Examples of such parameters are the placement, duration, extension and the speed with which a behaviour is executed. Head nods, for instance, can differ in the speed, extension and number of repetitions. One nod could be a small nod, performed very slowly while another nod could be very fast and aggressive. This variation might correspond to a systematic difference in meaning.

The classification of a signal based on the settings of the parameters along which it can vary is similar to the disambiguation of a signal in context; just as a head nod can mean different things in different contexts, a nod can also mean different things when the parameters are different. The question may thus arise whether a variation in context should not be seen as a variation within the signal. To clarify this, a short example will be given. Image that a person is

saying 'I mean that car.' and he wants to emphasize the word 'that'. He does this by giving a small nod and raising his eyebrows at that moment. The question is what should be seen as a signal and what as context. The head nod could be a single signal, and the eyebrow raise and the word 'that' as the context. But the head nod plus the eyebrow raise could also be seen as a single signal (an accent) with only the utterance as the context. Or in the extreme case the complete setting (head nod plus eyebrow raise plus 'that') could be seen as one single signal with no other signals as context. However, the result for the disambiguation of the signals would be the same. In the first case the intended meaning has to be found of a nod as the eyebrow raise and the utterance as its context. In the second case the signal is more extensive (the head nod and the eyebrow raise) resulting in fewer possible functions, but as the context is also smaller the result is the same. In the last case there is no context to use for disambiguation, but the signal is so specific (head nod plus eyebrow raise plus 'that') that there are probably not a lot of possible intended functions.

What this example shows is that for signal disambiguation it does not really matter where the segmentation line is drawn. What matters for disambiguation is the complete set of data, the information of the signal itself (the parameters) plus the context. It does not matter where the line between these two is drawn because the complete set stays the same. However, a line must be drawn to define what a signal is exactly. As explained, the parameters of that signal are just as important as the context, thus, the first group of contextual elements consists of the *parameters* of the signal itself. The way signals are described determines how they should be mapped to conversational functions.

3.2 What Does the Context Change

Now it is clear that before trying to disambiguate signals they should be concretely defined first. Only then is it possible to look at the context. In a conversation this context influences several things, but what exactly is the effect of the context on the behaviour of people? First of all the context plays an important role in determining the actual multi-modal behaviour that is used to express a conversational function. Maybe several modalities are already occupied (one has just taken a bite of sandwich), or maybe the situation asks for a certain modality rather than another (for example gesturing in a crowded environment).

Going one step higher in the process the context also modifies the function choice. The complete context (previous behaviour, speaker/listener role of the participants, etc) determines what conversational functions the agent or human can or cannot express. The contextual elements that do this are *constraining elements*; they constrain the function choice by defining what is and, mostly, what is not possible in the current context. One type of constraint is an expressive constraint which make it simply impossible to express a certain function. For example, a person cannot decline a turn if this was not offered to him first. And a person that tries to give the turn away when he does not have it will end up showing a continuation signal, signalling to the speaker to continue speaking, instead of a give turn signal. Another type of constraint is an appropriateness

constraint, meaning that certain functions just do not fit in the situation. An example would be to give the turn during a sentence while continuing speaking, or greeting a person when you are already talking for quite a while. All these actions are not appropriate or possible in certain contexts and are therefore not expected in those contexts.

When trying to disambiguate a signal, the first kind of modification behaviour of context (influencing the choice of behaviour to express a function) is not relevant. It does not really matter which signal was chosen to express a certain function, the important part is to find the meaning behind the signals. The second modification behaviour however defines the actual problem. One has to know the context to know which functions are appropriate at a certain point and this knowledge can be used to determine what was intended with a detected signal. In this case, the context can act as a filter, making certain interpretations unlikely.

3.3 Ambiguity

The previous paragraph characterized disambiguation of signals as a process of using the context to determine what conversational functions are appropriate at that time. This means that it can happen that the number of possible (appropriate) functions (based on the context) is greater than one. To solve this problem the context has to be used to find *pointer elements*. These elements do not constrain the function choice, but they point in the right direction. They provide the function that was most likely meant. For example, a head nod can mean different things: it can mean yes, it can be a backchannel, it can express an intensification (in the case of beats), etc. If a nod is shown just after a direct yes/no question then the nod is probably meaning yes. But if the nod is expressed at the same time that 'uhuh' was uttered then it was probably a backchannel. Of course there are always exceptions, but these pointer elements from the context can help picking out the most likely function.

Of course it can still happen that there are two or more best guesses, in which case not only the signal is ambiguous (they usually are, which is why disambiguation is needed), but the signal plus the context is ambiguous. So even when taking the complete context into account the signal can still be interpreted as multiple intentions.

This can mean two things. It is possible that only one function is intended by the signal but the signal was 'badly' chosen, meaning that that particular signal was not clear in that context. Thus, for the receiver it is not possible to decide what was intended and an educated guess needs to be made. If the wrong guess is made, the conversation might break down, or a repair action is taken.

It is also possible that the person or agent producing an ambiguous signal intended to communicate all the different meanings. Take for example a backchannel utterance. According to [12] this can be a continuer, an indicator of the listener that he does not want the turn and that the speaker should continue speaking. Another function that uses a backchannel utterance as a signal is an acknowledgement [2,4,6]. In a lot of contexts the difference between these two functions is hardly visible, in a lot of situation both a continuer and an

acknowledgement would fit. But it is not unimaginable that a backchannel utterance means both at the same time, which means that the actor really is saying "I'm still following you, please go on".

When disambiguating signals, these types of ambiguities should be kept in mind and it should be realized that sometimes the function of a signal simply is not clear (of course this is not the desired situation but it might happen in certain cases), and sometimes a signal can have multiple meanings at the same time.

3.4 Contextual Elements

An important task in signal disambiguation is to determine what elements from the context are important. For example, it is important to know whether the actor of a signal is the listener or the speaker. Other signals that are sent at the same time are important as well, as is the placement of the signal in the sentence (middle, end). A full list of these contextual elements will not be given here, but based on the previous sections (in which they are already named) they can be divided into three basic categories.

- *Parameters* – The first category of contextual elements contains the parameters of the signals themselves. The way signals are described determines how they should be mapped to conversational functions. When disambiguating signals, the first thing that has to be done is specifying exactly what defines a signal and to what conversational functions these signals (without taking context into account yet) can be mapped. For example, a list has to be created of all possible intentions of an eyebrow raising and an eyebrow lowering.
- *Constraining elements* – The second category contains the elements that constrain the presence of functions in certain contexts. They can be expressive constraints (some functions are impossible to express in certain contexts), or appropriateness constraints (some functions are very inappropriate in certain contexts). For example, a person can not decline a turn he does not have (expressive constraint) and it is very inappropriate to greet a person when you are already talking for a long time (appropriate constraint).
- *Pointer elements* – The last category contains the elements that do not restrict functions, but help to find the most likely one. For example, a nod after a yes/no question probably be interpreted as 'yes', while a nod together with a 'uhuh' probably is a backchannel.

4 Conclusions

In this article the disambiguation problem, mapping detected multi-modal behaviour to intended communicative functions was discussed. To solve this problem the context of the behaviour has to be used and this context also contains the parameters of the signals you are detecting. Using this information the task is to make a list of all communicative functions that are possible in the current context, merge this with the list of functions that the target behaviour can mean

and use the resulting as the intended function. Or, if the list of possible functions in the context is too large (maybe even infinite) you can do it the other way around by checking what a signal can mean and then check which meaning fits the current context.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE).

References

1. Beavin Bavelas, J., Coates, L., Johnson, T.: Listener responses as a collaborative process: The role of gaze. Journal of Communication 52, 566–580 (2002)
2. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit, Pittsburgh, Pennsylvania, United States, pp. 520–527. ACM, New York (1999)
3. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F. (eds.): Embodied Conversational Agents. MIT Press, Cambridge (2000)
4. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
5. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding, semiotical. Semiotica 1, 49–98 (1969)
6. Ekman, P.: About brows: Emotional and conversational signals, pp. 169–202. Cambridge University Press, Cambridge (1979)
7. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman Pub. Group (May 1976); published: Paperback
8. Heylen, D.: Challenges ahead: Head movements and other social acts in conversations. In: Halle, L., Wallis, P., Woods, S., Marsella, S., Pelachaud, C., Heylen, D. (eds.) AISB 2005 Social Intelligence and Interaction in Animals, Robots and Agents, pp. 45–52. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Hatfield (2005)
9. Lyons, J.: Introduction to Theoretical Linguistics. University Press, Cambridge (1968)
10. Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J.: Towards a model of face-to-face grounding. In: ACL 2003: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, vol. 1, pp. 553–561. Association for Computational Linguistics (2003)
11. Samtani, P., Valente, A., Johnson, W.L.: Applying the saiba framework to the tactical language and culture training system. In: Padgham, Parkes, Müller, Parsons (eds.) 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal (May 2008)
12. ten Bosch, L., Oostdijk, N., de Ruiter, J.P.: Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS, vol. 3206, pp. 563–570. Springer, Heidelberg (2004)