Selecting Appropriate Agent Responses based on Non-Content Features

Mark ter Maat
Human Media Interaction, University of Twente
PO Box 217, 7500 AE Enschede, the Netherlands
maatm@ewi.utwente.nl

Dirk Heylen
Human Media Interaction, University of Twente
PO Box 217, 7500 AE Enschede, the Netherlands
heylen@ewi.utwente.nl

ABSTRACT

This paper describes work-in-progress on a study to create models of responses of virtual agents that are selected only based on non-content features, such as prosody and facial expressions. From a corpus of human-human interactions, in which one person was playing the part of an agent and the second person a user, we extracted the turns of the user and gave these to annotators. The annotators had to select utterances from a list of phrases in the repertoire of our agent that would be a good response to the user utterance. The corpus is used to train response selection models based on automatically extracted features and on human annotations of the user-turns.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Discourse

General Terms

Human Factors

Keywords

Virtual Agents, Machine Learning, Behaviour Selection

1. INTRODUCTION

The key to a good conversation is content; knowing what the conversation is about and giving a good informational response. At least, that is what is often crucial in most practical spoken dialogue systems. And since it is not possible to include all 'world knowledge' in a system, most spoken dialogue systems stick to a very small domain and a simple task.

The Semaine Project is different. In this project, we try to create a SAL, a Sensitive Artificial Listener. This is an ECA — an Embodied Conversational Agent — that can react appropriately as a listener to the user's speaking behaviour, and that can motivate the user to keep on speaking.

AFFINE'10, October 29, 2010, Firenze, Italy.

One of the challenges in this project is to give appropriate feedback without knowing anything about the content of the conversation, which also means that the user is not bound to a particular domain or topic. As an extra challenge, the system consists of four characters with different personalities (corresponding to the four extremes of the valence and arousal model of emotion), and each character has the additional goal of getting the user in its own emotional state.

The ultimate goal is to make the virtual agent domain-independent. However, even though there is no fixed domain in the Semaine project, the questions asked by the virtual character are all about the life of the person that is speaking with the agent. This results in the users speaking about their work, their hobbies, holidays, and events that happened to them in the past. The virtual character responds to this with a response from a fixed set (about 100 per character). Using a Wizard-of-Oz setup, Douglas-Cowie et al. [3] demonstrate that a limited set of responses can be enough to sustain a conversation for some time, sometimes in the order of half an hour.

Instead of looking at what the user is saying, we focus on all the other verbal and nonverbal data that is available. For example, from the speech the energy and prosodic information can be extracted, together with a small set of keywords [12]. These keywords give an indication of what the user is speaking about, for example about the past or the future, about him or herself, or about other people. Based on all these features, some higher-level features can be extracted, for example valence and arousal [9], interest, and emotions [5].

The user's head movements also are a valuable source of information. Head movements can be detected, which can be used to detect nods and shakes. From the face, facial action units can be detected and tracked, which provides information about movements in the mouth, the eyes and the eyebrows [7]. This information leads to higher-level features too, such as agreement [1] and emotions (for example happy or sad) [11].

However, one of the key issues in the Semaine project is what the virtual characters should say, even though their repertoire is limited. How can the agent choose a suitable response, based on all the non-content features it receives? This paper describes the experiment to find response selection rules based on studying human-human data. Section 2 explains how the system currently chooses its responses, and how this could be improved. Section 3 shows how data was gathered, and section 4 shows some initial results. Section 5 looks forward, and how to continue from where we are now.

Cluster 1	Cluster 2	Cluster 3
Thats awfully sad,	Tell me about the last time you were really hurt,	Im not sure that you should be so positive,
Its a hard row to hoe,	Tell me about the things that get you down,	Im not sure that you should be so pragmatic,
It can be so depressing,	Tell me what gets you down,	Dont know what youve got to feel so cheery about,
Theres not much you can do about it,	Tell me about the things you dont want to think about,	If youre laughing, people wont take you seriously,

Table 1: Some utterances from three clusters of Obadiah's data

2. RESPONSE MODELS

In the first part of the project, no annotated data of SAL-like conversations were available. Therefore, the first versions of the agent were build using handcrafted rules. The Dialogue Manager consists of a collection of modules that each implement rules of a certain type. These simple rules check for certain features in the previous user utterance, and select a certain response. When selecting an utterance, the system asks each module to return a set of possible responses with a value indicating the quality of each response. Then the dialogue manager will select the response with the highest quality value. As an example, one of the rules, the *Content* rule, was based on the analysis of the SAL dialogues produced within the HUMAINE project [4]. Based on keyword spotting, abstract category labels were attached to the user utterances, for example about the time (past, present, or future), who the user is speaking about (him or herself, the agent, other people), and the polarity or semantic orientation of the utterance (positive or negative). Based on the HUMAINE data, a mapping was created from these labels to a response of the agent.

Another example is the *Arousal* rule, which is based on a major premise of the Semaine dialogues: characters try to influence the arousal and valence levels of the interlocutors. The arousal-rule thus specifies for each character how to respond to low and high values of detected arousal of the user. These rules were therefore not totally ad hoc, but certainly preliminary, and better rules — based on data rather than rule of thumb — are needed. Also, the current rules do not display the right amount of sensitivity to the user and thus fail to satisfy one of the major goals of the project: to create *Sensitive* Artificial Listeners.

Instead of handcrafting the rules, a better approach would be to extract them from data. For this we need a set of user utterances, and for each utterance a possible, appropriate response. Using statistics and machine learning techniques it is then possible to extract rules, which map certain features to certain responses.

Of course those user utterances cannot be used directly: features need to be extracted. The most practical choice would be to use automatically detected features, since the system can use the exact same features. This results in very applicable models, which can be used directly in the system. The downside is that these features are not robust, and contain a lot of mistakes, especially with the more complex features such as emotions. Another interesting choice is to use human-made annotations of the user utterances. This results in models that provide a lot of useful information, for example which annotations are useful, but the models cannot be plugged directly into the system. On the other side, these features are very reliable and informative, since humans can 'detect' a lot of subtle differences.

Since both resulting models are useful, we decided to do both. We can use the annotation-models to learn and to get a better understanding of the context, and the detection-models to plug into the system.

The next section describes the data we used.

3. DATA

Recently, the Semaine data was released [10]. This corpus contains 100 conversations with 20 participants, in which one participant plays the role of a listening agent, while the other participant is the speaker. During the conversations, the participants were seated in separate rooms, seeing each other through teleprompter screens. The conversations were recorded by five cameras and 4 microphones. Each conversation lasts approximately five minutes.

All recorded conversations are fully transcribed, and every conversation is annotated for five affective dimensions: Valence, Activation, Power, Anticipation/Expectation and Intensity. Next to this, an additional 25 dimensions are partly annotated. Among these dimensions are the six basic emotions (such as happiness and sadness), epistemic states (such as agreement, interest, and thoughtfulness), interaction process analysis components (such as solidarity, tension, and asking and giving opinions and information), and validity components (such as social concealment, and breakdown of engagement).

Using the audio feature extractor which we use in our project, OpenSMILE (the core component of OpenEAR [5]), we extracted audio features from the Semaine recordings. It extracts low-level features — such as the F0-frequency, the pitch direction, and the energy — and higher-level features — such as valence, arousal, and interest.

However, there is a big problem with the conversations. The SAL agent can select its response from a finite list of responses — about 100 for each character. But the participants in the listener role were not restricted by this list: they could say whatever they wanted. Of course this makes it impossible to extract response selection rules directly from the data.

In order to use the data we took out all the user-turns (a period of speech of the user between two listener responses) and showed them to human annotators. These annotators then had to select three possible responses from the list of agent responses. This means watching and listening to the user-turn and determining what SAL response would be appropriate after that user-turn. At the moment, 519 user-turns have been annotated by two annotators, with about 200 more user-turns on the way.

	C4.5	RIPPER	Multilayer Perceptron
Start new subject	0,57	0,57	0,67
Respond to good news	0,64	0,57	0,69
Respond to bad news	0,58	0,36	0,58
Ask for more info	0,43	0,46	0,47

Table 2: Precision results for three machine learning techniques with all labels separated.

	C4.5	RIPPER	Multilayer Perceptron
Other three combined	0,68	0,66	0,64
Respond to good news	0,65	0,67	0,64

Table 3: Precision results for three machine learning techniques trained on 'Respond to good news'.

4. INITIAL RESULTS

This section describes the initial results of this study. It describes how the data was preprocessed, the machine learning and statistical techniques that were used, and some results that were retrieved.

4.1 Preprocessing

The first thing to do is preprocess the data, that is, getting a table of meaningful features of the user-turns, and for each set of features a meaningful label. Currently, only the automatically detected features are used, and, as mentioned in the previous section, these contain low-level features and higher-level features. OpenSMILE sends the low-level features every 10 ms, which is unusable because the feature-set should have a fixed length. Therefore, of every low-level feature the minimum, the maximum, and the average value was calculated. The higher-level features are event-based, and consist of a value between -1 and 1. These events are summarized by the following features: number of low (< 0), number of high (> 0), the minimum, the maximum, and the average value. After this, the uninformative features were filtered out. For example, the minimum F0 value was always zero, and the feature voice_probability gives the chance that the user is speaking, which is also useless because the data is about user-turns where the user is speaking.

The next task is to assess the labels (the annotated responses). The problem with these labels is that there are too many possible responses, that is, more than 100 per character. Thus, we have to group the responses, since a lot of responses have the same meaning or intention, and can therefore be grouped together. This will significantly decrease the size of the label set. However, determining which responses to group is not trivial. Grouping them ourselves is very subjective, and might lead to poor results. A better approach is to base the groups on the data itself.

Using the expectation-maximization (EM) algorithm we grouped together the user-turns of which the data points were very close together. Then we looked at the responses belonging to those user-turns, and use these groups to cluster the responses. An example of such a clustering can be found in Table 1. This table shows three clusters with some responses of Obadiah, the sad character with low arousal and low valence.

As can be seen, these responses intuitively belong in the same cluster. And since this clustering is based on the actual data, this is a very promising result. We found similar results for the other characters. When trying to name the clusters, we found the following four types of clusters with each character: 'Start new subject', 'Ask for more info', 'Respond to good news', and 'Respond to bad news'.

4.2 Machine learning

With the data preprocessed and formatted, it is time to start the machine learning. Since the goal of this study is to find response selection rules, we need algorithms that do not just work, they need to specify how they work too so we can extract rules from them. For that reason, we chose a decision tree algorithm, J48, which is a Java implementation of the C4.5 algorithm [8], and we used RIPPER, an algorithm that tries to find rules [2]. To check whether a 'black box' machine learning algorithm (one that does not produce rules) performs better, we also decided to use a Multilayer Perceptron [6].

Since the Semaine system contains multiple response selection modules, it is no problem if a certain module produces zero suggestions on some occasions. What is important is that the responses it produces are appropriate: it is better to respond with a general response such as 'Tell me more' when the agent is not sure, than to respond with a specific response (for example 'I think you've done very well') in the wrong context. Because of this, the precision of the models is the most important property, we want this to be as high as possible.

4.3 Some results

Table 2 shows the precision-values of the three machine learning techniques that were mentioned. In this trial we ordered the classifiers to distinguish between the four different labels of Poppy, the positive and active character. In a second trial we wanted to see whether the results would improve if we would classify on a single label (grouping the other three labels together). The results of this for the label 'Respond to good news' can be seen in Table 3.

When classifying for all four labels, the Multilayer Perceptron clearly performs better than the other two algorithms. However, as we said before, this technique does not produce any rules we can use. Of the rule-based techniques, the C4.5 algorithm performs best. However, the decision tree this algorithm produced contains 23 nodes (decision points) and 24 leaves. These are a lot of rules, and the risk exists that this model was over-classified for the data. The Ripper algorithm performed better in that sense: it produced only nine rules, for example this one:

IF logEnergy_avg <= -16.52595, AND
rmsEnergy_avg <= 0.0017
THEN response = Start new subject

When classifying only two labels ('Respond to good news', and the other three combined), the Multilayer Perceptron suddenly performs worst. The C4.5 and the Ripper algorithm perform about the same, but both algorithms can recognize the label 'Respond to good news' better than when classifying four labels. This time, the decision tree contains 19 decision nodes and 20 leaves, but the Ripper algorithm produced only one rule:

IF interest_avg > -0.767
THEN response = Respond to good news
ELSE response = Other

5. DISCUSSION AND CONCLUSION

Instead of creating dialogue rules for very specific domains based on content, the Semaine project tries to create rules that are independent of the domain, and are based on non-content features such as prosody, a small set of keywords, head movements and facial expressions. The upside of this approach is that when extending the system, it does not have to be able to detect and understand more and more words. Also, the responses are independent of the current topic, which means that the agent is able to say something useful in a lot of situations.

Based on the non-content features, it is possible to respond with an appropriate response that is more than a simple 'Tell me more'. A lot of information can be extracted from these features, such as the emotional state of the users, their affective state, and whether they agree or disagree. This information can be used to select an appropriate response.

The initial results look very promising. The different algorithms produce rules that can predict an appropriate response-group with a precision between 0.43 and 0.69, varying per response-group. However, a lot more work is needed. More data is coming in, which should increase the robustness of the models. With all new data, rules have to be found for each response-group and for each character. However, when a model predicts a certain response that is not annotated, it does not automatically mean that it is not appropriate. Therefore, in order to evaluate the models, these models should produce responses for new data. These responses should then be evaluated by humans. Only then is it possible to say something about the real performance of the models.

6. ACKNOWLEDGMENTS

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).

7. REFERENCES

[1] K. Bousmalis, M. Mehu, and M. Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *Proceedings of IEEE Int'l Conf. Affective Computing and Intelligent Interfaces (ACII'09), Amsterdam*, volume 2, pages 1–9, Los Alamitos, September 2009. IEEE Computer Society Press.

[2] W. W. Cohen. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufman, 1995.

[3] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. K. J. Heylen. The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. In L. Devillers, J.-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4, Paris, France, 2008. ELRA.

[4] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 488–500, Berlin, Heidelberg, 2007. Springer-Verlag.

[5] F. Eyben, M. Wöllmer, and B. Schuller. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, pages 576–581, Amsterdam, The Netherlands, 2009. IEEE.

[6] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.

[7] S. Koelstra and M. Pantic. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. In *Proceedings of IEEE Int'l Conf. Automatic Face and Gesture Recognition (FG'08)*, pages 1–8, Amsterdam, The Netherlands, 2008. IEEE.

[8] S. L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, 16(3):235–240, september 1994.

[9] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Merano, Italy, 2009. IEEE.

[10] M. Valstar, G. McKeown, R. Cowie, and M. Pantic. The Semaine Corpus of Emotionally Coloured Character Interactions. In *Proc. IEEE Int. Conf. on Multimedia & Expo (ICME2010)*, Singapore, 2010. IEEE.

[11] M. Valstar and M. Pantic. Biologically vs. logic inspired encoding of facial actions and emotions in video. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 325 –328, 9-12 2006.

[12] M. Wollmer, F. Eyben, B. Schuller, and G. Rigoll. Robust vocabulary independent keyword spotting with graphical models. In *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 349 –353, nov. 2009.