Extracting Emotional Information from the Text of Spoken Dialog

Curry Guinn and Rob Hubal
RTI International, 3040 Cornwallis Road,
Research Triangle Park, North Carolina, USA, 27709
{cig, rhubal}@rti.org
http://www.rvht.info

Abstract

This paper describes research in developing tagged semantic grammars that carry emotional and attitudinal information about the user's utterance. This information is then used to characterize the emotional state of the user in its interaction with a virtual computer characters. This paper describes several applications that use tagged semantic grammars and presents some results from those systems.

1. Introduction

This paper describes research in developing tagged semantic grammars that carry emotional and attitudinal information about the user's utterance. Certain fragments of an utterance may be labelled with emotional and attitudinal content, and these values are combined as the utterance is completely parsed. In addition to the semantic content of the utterance, emotional and attitudinal information is passed to the dialog manager which utilizes this information to modify its model of the user. This methodology has been employed in a number of applications including systems for training police officers for encounters with the mentally ill (JUST-TALK) and for training telephone survey interviewers (AVATALK-Survey).

1.1 Paper Outline

The structure of a semantic grammar and how it is tagged is presented in Section 2. The mechanism for combining the emotional/attitudinal tags obtained from each sentence fragment in a full utterance parse is presented in Section 3. The overall system architecture in which these tags were used is presented in Section 4. Section 5 discusses two applications using this methodology and presents some results from user studies.

2. Emotionally Tagged Semantic Grammars

Semantic grammars are a very common form of language representation for spoken natural language processing system. These typically domain dependent grammars directly map the incoming text (typed or output from speech recognizer) directly to an underlying semantics.

2.1 Semantic Grammars for Parsing

An example of a fragment of a semantic grammar is presented in Figure 1 from the domain of asking a location. This example grammar is context-free with a single, unrestricted non-terminal on the left hand side. Notice that some rules have semantic information that follows the colon (':'); this information is what gets returned by the parser to the dialog manager.

```
S -> ASK LOC' : ask(location(LOC')) .
ASK -> PLEASE WHEREIS .
ASK -> WHEREIS .
PLEASE -> damn it .
PLEASE -> please .
PLEASE -> would you please .
WHEREIS -> help me FIND .
WHEREIS -> where are .
FIND -> find .
FIND -> locate .
LOC -> my KEYTYPE' keys : KEYTYPE' .
LOC -> my shoes : shoes .
KEYTYPE -> house : house .
KEYTYPE -> car : car .
```

Figure 1. Fragment of a Semantic Grammar

2.2 Applying Emotional/Attitudinal Tags to Grammar Rules

The above grammar may be augmented by attaching emotional/attitudinal tags to each grammar rule. For instance, the designer of the grammar may decide that the use of the word "please" adds to the politeness of the sentence. Thus the rule

```
PLEASE -> please POLITENESS 0.2 .
```

would indicate that use of the rule in parsing the phrase would increase the overall sentence politeness by a small amount. Values between -1.0 and 1.0 are assigned to emotional tags. Thus a value of 1.0 for POLITENESS would be the maximum value for politeness, while -1.0 would be the most impolite phrase. The grammar presented in Figure 1 might be modified as is shown in Figure 2 to indicate levels of politeness.

```
S -> ASK LOC' : ask(location(LOC')) .
ASK -> PLEASE WHEREIS .
ASK -> WHEREIS .
PLEASE -> damn it POLITENESS –0.4 .
PLEASE -> please POLITENESS 0.2 .
PLEASE -> would you please POLITENESS 0.3 .
WHEREIS -> help me FIND POLITENESS 0.1 .
WHEREIS -> where are .
FIND -> find .
FIND -> locate .
LOC -> my KEYTYPE' keys : KEYTYPE' .
LOC -> my shoes : shoes .
KEYTYPE -> house : house .
KEYTYPE -> car : car .
```

Figure 2. Modified Grammar with Politeness Tags

2.3 Possible Tags

In practice, a number of tags proved identifiable and useful in the application domains. Politeness, urgency, satisfaction, anger, confusion, complexity, and formality were some of the most commonly used. These tags would be put in by hand as the grammars were built or inserted afterwards.

In our system, multiple coders assign values to particular grammar fragments. Variations and discrepancies are mediated. A more scientific approach for determining appropriate values could be obtained by obtaining relative rankings of the emotional impact of words and phrases across a wider range of human subjects. Several studies have been conducted in this area (Pennebaker et al, 2003 gives an overview of this research). However, analysis in this area tends to be focused on particular emotions (like politeness in Brown & Levinson, 1987). As we attempt to build more realistic behaving agents, continued research in this area is needed.

2.4 A Word on Minimum Distance Parsing

The parsing technology employed relies on a minimum distance translator (MDT) algorithm [Hipp, 1994]. In an MDT, upon receipt of an utterance, the algorithm translates the utterance to its closest match within the allowed utterances defined by the semantic grammars. This parser was augmented to be able to process the emotional tags. One advantage of the MDT algorithm is that allows for ill-formed input, that is, input that does not exactly match anything in the grammars. The algorithm attempts to find the best match for something in the grammar and returns a score indicating the quality of the match.

3. Combining Emotional/Attitudinal Tags From Grammar Fragments

In some instances a single utterance may be parsed using multiple rules, some of which may contain the same emotional or attitudinal tag. In the grammar given in Figure 2, the input sentence "Please help me find my car keys." would be most accurately parsed with the rules:

```
S -> ASK LOC' : ask(location(LOC')) .
ASK -> PLEASE WHEREIS .
PLEASE -> please POLITENESS 0.2 .
WHEREIS -> help me FIND POLITENESS 0.1 .
FIND -> find .
LOC -> my KEYTYPE' keys : KEYTYPE' .
KEYTYPE -> car : car.
```

Two of those rules indicate a modification in the level of politeness. How do these values get combined coherently? Simply adding and subtracting the values will not necessarily keep the values between -1.0 and 1.0. Instead, the algorithm in Formula 1 is applied. This combination rule is similar to what is used to combine independent variables in probability. For each emotional or attitudinal tag *i*, the values are combined in the following way:

$$1 - \prod (1 - Tag_i) \qquad (1)$$

In the example given above, the POLITENESS values are 0.2 and 0.1. Thus the combined values would be 1 - (1 - 0.2)(1 - 0.1) = 1 - (0.8)(0.9) = 1 - (0.72) = 0.28.

3.1 Normalization

For readability and for ease of coding, emotional and attitudinal values are coded between -1.0 and 1.0. Intuitively it is much easier to understand the politeness of "You idiot!" as a negative value rather than just a value less than 0.5. For computation, however, the values are normalized so that they are between 0.0 and 1.0. Thus a value of POLITENESS of 0.0 becomes a normalized value of 0.5, a POLITENESS of -1.0, becomes 0.0, and POLITENESS of 1.0 is normalized to 1.0.

4. Integration

We have developed a PC-based architecture, Avatalk, that enables users to engage in unscripted conversations with virtual humans and to see and hear their realistic responses [Guinn and Montoya, 1998, Hubal and Frank, 2001]. Among the components that underlie Avatalk are a Language Processor, a Behavior Engine, and a Visualization Engine (Figure 3).

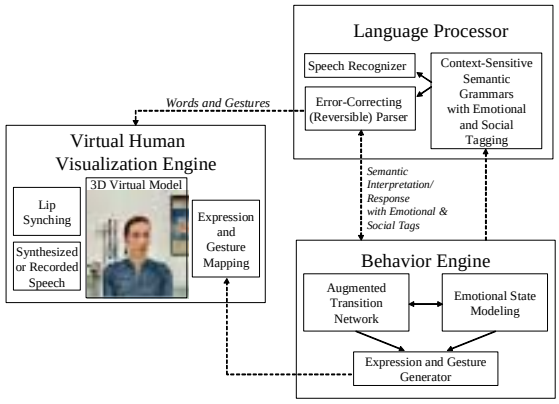

Figure 3. Avatalk Architecture

The Language Processor accepts spoken input and maps this input to an underlying semantic representation, and then functions as a speech generator by working in reverse, mapping semantic representations to speech output, facial expressions, and gestures, displayed by the Visualization Engine.

The Behavior Engine maps the output of the Language Processor and other environmental stimuli to agent behaviors. These behaviors include decision making and problem solving, performing actions in the virtual world, changes in facial and body expression (via the Visualization Engine), and spoken dialog. In the example given above, the Behavior Engine would determine how the virtual character should respond, given the semantics and the calculation that the input was reasonably polite. The Behavior Engine also controls the dynamic loading of contexts and knowledge for use by the Language Processor. Within the Behavior Engine is a model of the user's emotional state. These values are updated based on the input from the emotional and attitudinal tags from the Language Processor and inferences made from the actions the user has taken.

The user model is based on prior art [for details see Hubal, Frank, and Guinn, 2003]. Expert-derived transition tables guide emotional state updates. Each application employs different combinations of tagged emotions and attitudes, appropriate for the user model and utterance effects on the virtual character.

The Visualization Engine takes gesture, movement, and speech output and enables the 3D representation of a human to perform these actions. It accomplishes these movements through morphing of vertices of a 3D model and playing of key-framed animation files (largely based on motion capture data). The Visualization Engine is capable of lip-synching to both synthesized and recorded speech.

5. Applications

The Avatalk architecture has been employed in a number of applications. Two applications which have been used in classroom training environments include JUST-TALK, a law enforcement trainer for managing encounters with the mentally ill, and AVATALK-Survey, a telephone survey interviewer trainer.

5.1 JUST-TALK

JUST-TALK teaches students basic techniques for managing encounters with the mentally ill by having them work through a series of one-on-one scenarios with a simulated subject [Frank et al, 2002]. It also teaches them to look for indications of particular forms of mental illness so that they can adapt their responses appropriately. Through observations of the virtual environment and a dialog with the virtual subject, the student must stabilize the situation and decide whether to release or detain the subject.

Based on expert advice, user input for JUST-TALK is assessed for politeness, personalization, and complexity, in addition to higher-level analyses of syntactic form and responsiveness. For instance, users are expected to be polite; absence of "sir" or "madam" leads the virtual character to become more wary or agitated. Similarly, complex sentences for certain characters (depending on initial mental state) increases confusion or fear.

JUST-TALK was delivered to the North Carolina Justice Academy where it has been used in three courses with a total of 44 students. Each three-day class for law enforcement personnel on handling encounters with the mentally ill included classroom lecture, videos, live role-plays, and simulated role-playing using the JUST-TALK software. Students filled out an evaluation of the system after its usage. One of the most encouraging results was that 56% of the students found the software as useful or more useful than live role-playing.

5.2 AVATALK-Survey

The same Avatalk architecture was employed for training telephone survey interviewers in the process of obtaining cooperation [Link et al, 2002]. Using a voice interface alone (no 3D visualization) the trainees interacted with the virtual humans as if they (the virtual humans) were the subjects of a phone interview. After using the training application for 20 or more dialogs, trainees were then asked to evaluate the system. 60% of the users rated the realism of the conversations to be somewhat to extremely high.

83% would recommend the application as a training tool for other interviewers.

6. Comparison to Past and Existing Work

6.1 Automated Information Extraction Tools

Existing e-mail and text messaging tools are intended to derive emotional and attitudinal content from text. For instance, MoodWatch is an e-mail feature that seeks potentially offensive text and tone in messages. Its developers understand that not all forms of flaming can be caught, yet by flagging occurrences of aggressive or demeaning language that fall into research-derived hierarchical patterns, a message writer can monitor his/her work that might unintentionally have been written in violation of "rules of common courtesy and social decorum" [Kaufer, 2000].

Other tools, such as MetaMarker [Magnenat-Thalmann and Kshirsagar, 2000], aim to determine the mood, content, and intent of text messages. After steps that include morphological, lexical, and syntactic analyses, discourse-level attributes are identified, including a measure of the emotional tone and urgency of the message. MetaMarker measures emotional intention as "strongly negative, negative, neutral, or positive" and urgency as "very urgent, urgent, and neutral".

These applications are most similar to the current work in their assessment of negatively valenced content. They are not, though, geared towards training, and they do not involve variable emotions or attitudes.

6.2 "Intelligent" Conversational Systems

The well-known (and badly misinterpreted) ELIZA [Weisenbaum, 1966] did not attempt to analyze emotional or attitudinal components of input. ELIZA simply maintained a conversation by reflecting the input, slightly modified, back to the user, according to a script of anticipated inputs and suitable (and non-repeated) responses. The more serious PARRY [Colby et al, 1972] generated different responses depending on what had happened prior to that input. PARRY interpreted input as neutral, questioning, sensitive, angry, or delusional, and responded normally, anxiously, hostilely, defensively, or evasively, according to pre-specified rules. The system reduces the input word by word to a root form that gets matched against a large database of anticipated inputs, each having values that affect internal emotional state variables. In effect, the PARRY approach most closely resembles that described in this paper, though only fear, anger, and mistrust are tracked.

Other researchers, similar to the system described in this paper, use "autonomous dialogue systems" [Magnenat-Thalmann and Kshirsagar, 2000] and "linguistic style" [Walker, Cahn, and Whittaker, 1997] to elicit emotional and attitudinal information from user input. Internal emotion or mood states and social interaction styles change based on emotionally tagged information or analysis of syntactic form and probability transition matrices. Primary differences include emotional/ attitudinal tags in an MDT parser, ease of substitution of different labels within the Language Processor grammars and Behavior Engine, and implementation for use in training applications.

7. Future Work

A challenge for our future research will be determine how much this mechanism affects a user's perception of the believability of the virtual characters. Given the presence of so many other confounding variables (facial gestures, body motions, character decision-making), it may be difficult to isolate the effect of this one mechanism. We use this mechanism both for detecting some of the user's emotion as well as for generating the virtual human's responses (grammars are reversible). It will be far easier to show the impact of the virtual humans use of word choice on user perceptions. The changes in the virtual character's behaviors based on the user's word choice tends to be far more subtle and has be inferred indirectly by the user.

While the performance of the system in these domains is satisfactory, a main concern is rapid expandability to other domains. The semantic grammars are somewhat modular with many components that can be reused. Nonetheless, a more thorough taxonomy of how grammars should be labelled with emotional and attitudinal content needs to be developed.

Parallel efforts are underway to augment emotional/attitudinal information with intonation, pitch, and other vocal input to better assess the emotional state of the user. The results of this analysis will directly integrate into the current architecture.

References

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge, England: Cambridge University Press.

Colby, K.M., Hilf, F.D., Weber, S., & Kraemer, H.C. (1972) Turing-like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes. Artificial Intelligence, 3, 199-222.

Frank, G., Guinn C, Hubal, R., Pope, P., Stanford, M., & Lamm-Weisel, D. (2002). JUST-TALK: An Application of Responsive Virtual Human Technology, Proceedings of the Interservice/Industry Training, Simulation and Education Conference.

Guinn, C.I., & Montoya, R.J. (1998). Natural Language Processing in Virtual Reality. Modern Simulation & Training, 6, 44-45.

Hipp, D.R. (1992). Design and development of spoken natural-language dialog parsing systems. Ph.D. thesis, Duke University, available as Technical Report, CS-1993-15.

Hubal, R.C., & Frank, G.A. (2001). Interactive Training Applications using Responsive Virtual Human Technology. Proceedings of the Interservice/Industry Training Systems and Education Conference.

Hubal, R.C., Frank, G.A., & Guinn, C.I. (2003). Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual Humans for Training. Proceedings of the Intelligent User Interface Conference.

Kaufer, D. (2000). Flaming: A White Paper. Found at http://www.eudora.com/presskit/pdf/Flaming_White_Paper.PDF.

Link, M.W., Armsby, P.P., Guinn, C., & Hubal, R. (2002). A Test of Responsive Virtual Human Technology as an Interviewer Skills Training Tool. Proceedings of the Annual Conference of the American Association for Public Opinion Research.

Magnenat-Thalmann, N., & Kshirsagar, S. (2000). Communicating with Autonomous Virtual Humans. Proceedings of the Twente Workshop on Language Technology.

Paik, W., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., & Amice, C. (2001). Applying Natural Language Processing (NLP) Based Metadata Extraction to Automatically Acquire User Preferences. Proceedings of the International Conference on Knowledge Capture.

Pennebaker, J, Mehl, M., and Niderhoffer (2003), Psychological Aspects of Natural Language Use: Our Words, Our Selves, *Annu. Rev. Psychol*., 2993, 54:547-77.

Walker, M., Cahn, J., & Whittaker, S. (1997). Improvising Linguistic Style: Social and Affective Bases for Agent Personality . Proceedings of the Conference on Autonomous Agents.

Weizenbaum, J. (1966). ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. Communications of the ACM, 9(1), 36-45.