Detection of Neuropsychiatric States of Interest in Text

Robert J. Bechtel

GB Software LLC
4607 Perham Road
Corona del Mar, CA 92625 USA
rbechtel@acm.org

Louis A. Gottschalk

Department of Psychiatry
University of California, Irvine
Irvine, CA 92697 USA
lgottsch@uci.edu

Abstract

This paper provides an overview of a technique for measuring neuropsychiatric states such as anxiety, hostility, and depression and of a software implementation of the technique. The implementation uses both clause structure and a dictionary of semantically-tagged words and phrases to assign scores to individual clauses and to aggregate those scores over larger samples. The basic technique and software implementation have been used in a variety of settings, and form the basis of current research toward an psychodynamic psychotherapist.

Obtaining Clinical Evaluations

Before a dialogue system can respond to affect, it must detect it. Psychiatry and psychology have long experience in the detection and measurement of affect, both in characterizing subjects, behaviors, and diagnoses, and in selecting and measuring the efficacy of interventions.

A clinician has several options in obtaining objective and valid clinical evaluations. For example, precision and accuracy may be avoided and impressionistic reactions and "gut feelings" can be relied on; some clinicians feel they are able to do competent clinical work with this approach. A clinician can spend considerable time and care in the diagnostic and therapeutic evaluation of children and adults with the goal of assessing accurately and precisely the magnitude of diverse psychopathological processes within patients at different times.

Another approach is to use various observer psychiatric rating scales, such as the Brief Psychiatric Rating Scale, the Hamilton Anxiety or Depression Rating scales or various self-report measures, such as, various adjective checklists. Although such measures are widely used in many research projects, their use carries with them a false sense of security since quite often no inter-rater reliability tests are done with the rating scales, the assumption being that anybody can follow the instructions for rating and no measurement errors are likely to occur. With rating scales, however, raters vary widely on how much of the range of ratings they use with the same subjects. Some raters characteristically select the lower range of the ratings; whereas others habitually chose the higher range of the ratings. With self-report measures, the assumption is that self-raters are all, indeed, in good and equivalent contact with themselves and are not likely to be falsifying, consciously or unconsciously, their self-evaluations, though it is true that the self-rating comes directly from the individual being evaluated.

These kinds of measurement errors in observer rating scales and self-report scales, usually disregarded by researchers and clinicians, are minimized in the measurement method of content analysis of verbal behavior. The subjects being rated are usually not aware what speech content or form is being analyzed, and they have difficulty covering up, even if they have some notions about such matters. Furthermore, the unstructured approach customarily used to elicit speech avoids the questionnaire or "prosecuting attorney" method, and allows the subject to elaborate and use free-will to the extent desired by the self on choice of topics to verbalize. Emotions, self-reflections, doubts, and defensive maneuvers are recorded, and these all contribute to the content analysis scores eventually calculated.

The content analysis approach to the measurement of psychological dimensions includes the strengths of both the self-report approach and the observer rating scale approach, and minimizes the weaknesses of both in terms of measurement errors. Of particular interesting in the current setting is that content analysis is particularly well-suited for use in dialogue systems.

The Gottschalk-Gleser Scales

While there are many content analysis scales and techniques, the Gottschalk-Gleser content analysis method (Gottschalk and Gleser, 1969) for measuring the magnitude of various psychobiological states and traits from the content analysis of verbal behavior has been successfully applied to many different neuropsychiatric dimensions. Extensive empirical research (Gottschalk *et al*, 1969) has established the validity and reliability of scales measuring a variety of emotional and psychobiological states including Anxiety, Hostility Outward, Hostility Inward, Ambivalent Hostility (hostility originating externally and directed towards the self), Social Alienation-Personal Disorganization, Cognitive Impairment, Hope, Depression, Human Relations,

Achievement Strivings, Dependency Strivings, and Health/Sickness.

Scores on the Gottschalk-Gleser content analysis scales are not simple word counts. The basis of analysis is the grammatical clause, requiring at least rudimentary syntactic analysis to determine clause boundaries. Assigning a tag to a clause may also require determining the agent and recipient of an action, and categorizing them as the self, other humans, subhuman, or inanimate. An example of a scale definition is given in Figure 1.

Figure 1: Anxiety Scale

1. Death anxiety -- references to death, dying, threat of death, or anxiety about death experienced by or occurring to:
   a. self (3).
   b. animate others (2).
   c. inanimate objects (1).
   d. denial of death anxiety (1).
2. Mutilation (castration) anxiety -- references to injury, tissue or physical damage, or anxiety about injury or threat of such experienced by or occurring to:
   a. self (3).
   b. animate others (2).
   c. inanimate objects destroyed (1).
   d. denial (1).
3. Separation anxiety -- references to desertion, abandonment, ostracism, loss of support, falling, loss of love or love object, or threat of such experienced by or occurring to:
   a. self (3).
   b. animate others (2).
   c. inanimate objects (1).
   d. denial (1).
4. Guilt anxiety -- references to adverse criticism, abuse, condemnation, moral disapproval, guilt, or threat of such experienced by:
   a. self (3).
   b. animate others (2).
   d. denial (1).
5. Shame anxiety -- references to ridicule, inadequacy, shame, embarrassment, humiliation, overexposure of deficiencies or private details, or threat of such experienced by:
   a. self (3).
   b. animate others (2).
   d. denial (1).
6. Diffuse or nonspecific anxiety -- references by word or phrase to anxiety and/or fear without distinguishing type or source of anxiety:
   a. self (3).
   b. animate others (2).
   d. denial (1).

While the utility of these scales has been demonstrated repeatedly through decades of research (Gottschalk, 1979, 1995), widespread, everyday use of content analysis of verbal behavior for research and clinical practice was hampered by the relatively high training and performance requirements associated with the manual application of the technique. For example, Gottschalk and Gleser (1969) recommend an inter-coder reliability coefficient of 0.80 or better with the scoring of qualified experts in the use of these content analysis Scales. To achieve this level of familiarity and skill in coding the scales requires some practice with previously published and unpublished examples of scoring these content analysis scales and continual monitoring of trained scorers. Manual scoring is also not a particularly quick process, requiring not only trained content judgments, but also extensive post-processing of scores to prepare scale-based summaries and analyses.

Computerizing the Scales

To address these training and performance obstacles to wider use of the scales, we have developed and tested software that is capable of reliably scoring computer-readable transcriptions of verbal (speech) samples on the Gottschalk-Gleser scales (Gottschalk and Bechtel, 1982, 1995). In operation, the program assigns scores on the user-selected scales to each clause in the input sample, then, at the user's option, reports score summaries for each scored scale with comparisons to established norms for the subject's demographic group, provides an analysis of the score profile, and suggests possible diagnoses drawn from the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) (American Psychiatric Association, 1994).

The software system relies on a reasonably-sized dictionary (~200,000 words) marked with associated parts of speech, and a collection of (mostly American) English idiomatic and slang expressions. Some of the words and all of the idioms have been manually identified as possible indicators of semantic content relevant to one or more of the scales. Syntactic information such as part of speech and number is extracted from the dictionary and used by a parser that outputs an analysis of the structure of each input clause.

When a word or phrase from the dictionary is noted as a possible marker of an item from a scale, it is added to a list of tagging candidates. This list of candidates is then examined by a set of scale-dependent procedures that consider the clause structure as well as the score-marking to decide the validity of each candidate. For example, the speaker must be the recipient of an action or affect described in a clause for the clause to be scorable on the Hostility Inward scale, while the speaker must not be the recipient on the Hostility Outward scale. Candidate tags approved by this process are emitted as content analysis tags applicable to the input clause.

A scoring summary is prepared for each active scale. The summary gives tallies of the number of occurrences of the

various tags, a word count, and a single score based on scale-specific tag weightings that is used to characterize the verbal sample on each scale. The summaries also indicate to what extent the verbal sample score deviates from the norms that have been obtained for each scale (in terms of standard deviations). These norms, available for children and adults, have also been derived from speech samples elicited by purposely ambiguous instructions requesting the adult speaker to talk for five minutes about interesting or dramatic personal life experiences. Finally, the system proposes possible neuropsychiatric diagnoses that the user might consider in evaluating the patient or subject. As noted earlier, the diagnoses suggested for consideration are taken from DSM-IV.

The software system does not always precisely match scores assigned by trained human scorers. Since the tagging dictionary was developed manually, occasional errors occur. Some tagging categories require judgments that we have been unable to capture in programmatic form. For example, on the Social Alienation/Personal Disorganization (Schizophrenic) scale, a clause is to be tagged if it contains "Obviously erroneous or fallacious remarks or conclusions; illogical or bizarre statements." To accommodate these differences, the score produced by the software is adjusted to create a "human equivalent" score, with the adjustment derived from a human and computer scoring of a set of 71 samples maintained as a standard.

Related Work

Not surprisingly, there is work by others that focuses on sensing and measuring emotion in text. In most cases, the goal is to enhance either computer-mediated human interaction or direct human-computer interaction. Liu, Lieberman, and Selker (2003) identify four common approaches (keyword spotting, lexical affinity, statistical methods, and hand-crafted models) before describing their knowledge-based multiple model technique. Madden (1999) provides a good example of keyword spotting, with keyword groupings for both emotion and personality. The fuzzy semantic typing approach used by Subasic and Huettner (2000) appears to fall within the lexical affinity approach as described by Liu *et al*. Guinn and Hubal (2003) use a semantic grammar, which could be viewed as a particularly sophisticated form of keyword spotting. Most similar to our approach is the Text-to-Emotion Engine of Zhe and Boucouvalas (2002), which uses keyword spotting coupled with structural constraints on parser output to restrict tag assignment.

All of these systems map into a standard set of emotion or affect categories. Liu *et al* and Madden use six emotional categories derived from the work of Ekman (1993). Subasic and Huettner use 83 affect categories, while Zhe and Boucouvalas developed a tagset with 119 categories, both apparently developed manually. Guinn and Hubal do not report the source or size of their tag, but do indicate that seven of their tags were most commonly used, suggesting a relatively small tag set. The Gottschalk-Gleser content analysis system has 14 distinct scales, with 278 tags within those scales. As noted earlier, the scales have been validated for construct validity and reliability (Gottschalk *et al*, 1969).

All these systems use some form of weighting as well as tag presence or absence. It appears that all mappings from text to emotion/affect tag are manual or heuristic at some level. That is, none of these systems have derived their tagging sets using clustering or other automated techniques. We are not aware of any study that compares these systems over a common set of inputs.

Interactive Application

While useful in a psychiatry or psychology research setting, for example in exploring the writings of the Unabomber for evidence of psychopathology (Gottschalk and Gottschalk, 1999), or in predicting the outcome of psychiatric commitment hearings (Lavid *et al*, 2002), our existing software is strongly oriented to the analysis of samples that are larger than many conversational utterances. For reliable results, the recommended minimum input sample length is 80 words The focus is usually on overall scale scores, aggregated over all the clauses in a sample, because at that level the established norms can be used for comparison purposes.

At least two efforts have been made to apply the scoring software in interaction settings. In the first (Gottschalk *et al*, 2003), the software was used to analyze utterances in dialogues among medical personnel and families of patients in a surgical intensive care unit. The problem of short utterances was addressed by aggregating utterances by speaker, enabling calculation of both a per-utterance and cumulative sample score for each speaker. The second interaction application is in a commercial web-based system that generates responses to subject entries, similar in spirit to the original Eliza (Weizenbaum, 1966), but with the (pre-written) response determined by the content analysis scores on the various scales (Journal Genie, 2004).

Roadmap

We are currently using our sample scoring software as a core for experimentation leading toward a capability for computer-based psychodynamic psychotherapy. In our experimental system, dialogue begins with a request that the participant summarize any complaints, worries, and symptoms. If the participant response is too brief for reliable scoring, simple "please tell me more" prompts are offered to elicit additional material. The goal is to obtain an initial understanding of the participant's mental status.

The participant response (possibly aggregated over several inputs) is analyzed and compared to the established norms for the various scales. If the scores assigned to the response deviate from the norms—that is, lie greater than one, two or three standard deviations from the established norm on any of the scales, the system shares the general analysis with the participant. On some occasions, the system may ask the participant for more information or to elaborate on some relevant theme in his/her initial verbal communication.

As an example, consider a participant response that is scored as greater than one standard deviation above the norm on the Social Alienation-Personal Disorganization scale and also greater than one standard deviation above the mean on the Anxiety scale. A system response would be something like "I gather from what you tell me that you have been experiencing some very uncomfortable mental symptoms. Can you tell me more about your self? And have you been getting some professional help for your condition?"

The affect detection and measurement capability of our existing sample scoring software supplies a useful starting point for our goal of therapeutic support, but it is only a small fraction of what would be needed to attempt psychotherapy. We are in the process of adding entity recognition and tracking to the core content analysis results to better establish the context in which affect appears. For example, if hostility outward is consistently detected with respect to a particular individual, interaction may be guided to examine that relationship in greater depth.

Much more work will be needed in a variety of areas before any meaningful attempt can be made at a sustained psychotherapeutic interaction. At a low level, system responses need greater flexibility than has been available in our simple text template approach. On the dialogue level, we need to incorporate the neuropsychiatric assessment within a user model, and create a model of the psychotherapy process that can be used to drive a dialogue move engine such as TrindiKit (Larsson and Traum, 2000). The need for interacting functionality at so many levels suggests that psychotherapy may offer a fertile ground for further work in dialogue research.

References

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual. Fourth Edition*. American Psychiatric Press.

Ekman, P. 1993. Facial expression of emotion. American Psychologist, 48:384-392.

Gottschalk LA. (Ed.) 1979. *The Content Analysis of Verbal Behavior. Further Studies.* New York: Spectrum Publications.

Gottschalk LA. 1995. *Content Analysis of Verbal Behavior. New Findings and Clinical Applications*. Hillsdale, New Jersey: Lawrence Erlbaum Publisher.

Gottschalk, LA and Bechtel RJ. 1982. The measurement of anxiety through the computer analysis of verbal samples. *Comprehensive Psychiatry*: 23:4 (July/August).

Gottschalk LA and Bechtel R. 1995. Computerized measurement of the content analysis of natural language for use in biomedical and neuropsychiatric research, Louis A. Gottschalk and Robert Bechtel, *Computer Methods and Programs in Biomedicine*: 47:123-130.

Gottschalk LA, Bechtel RJ, Buchman 2003. TA, Ray SE. Computerized content analysis of conversational interactions. *CIN: Computers, Informatics, Nursing*. 21:249-258.

Gottschalk LA and Gleser GC. 1969. *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. Berkeley, Los Angeles: California: The University of California Press.

Gottschalk LA and Gottschalk LH. 1999. Computerized content analysis of the Unabomber's writings. *American Journal of Forensic Psychiatry*. 20:5-31.

Gottschalk LA, Winget CN, Gleser GC. 1969. *Manual of Instructions for Using the Gottschalk-Gleser Content Analysis Scales: Anxiety, Hostility, Social Alienation-Personal Disorganization*. Los Angeles, Berkeley, University of California Press.

Guinn C and Hubal R 2003. Extracting emotional information from the text of spoken dialog. *Proceedings of the International Conference on User Modeling Workshop, Assessing and Adapting to User Attitudes and Affect: Why, When and How?*, June 22, 2003, Pittsburgh, PA.

Journal Genie website. 2004. http://www.journalgenie.com/

Larsson, S and Traum D. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. In *Natural Language Engineering* Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, pp. 323-340. Cambridge University Press, U.K.

Lavid N, Gottschalk LA, Grayden T, Bechtel RJ. 2002. Computerized content analysis of involuntary hospitalized psychiatric patients' requests to refuse hospitalization and medication. A preliminary study. *American Journal of Forensic Psychiatry*. 23:3, pp 55-69.

Liu, H, Lieberman, H, and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. *Proceedings of the 2003 International Conference on Intelligent User Interfaces*, IUI 2003, pp.125-132. January 12-15, 2003, Miami, FL.

Madden, S. 1999. WebEvaluator: deducing emotion and personality from web pages. http://www.cs.berkeley.edu/~madden/WebEvaluator/WebEvaluator.htm.

Subasic, P. 2000. Affect analysis of text using fuzzy semantic typing. *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, August 20-23, 2000

Weizenbaum J. 1966. ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9:36-45.

Zhe, Z and Boucouvalas, AC. 2002. Text-to-emotion engine for real time Internet communication." *International Symposium on Communication Systems, Networks and DSPs*, pp. 164-168. 15-17 July 2002, Staffordshire University, UK.