Characterizing Attitudinal Behaviors in On-Line Open-Sources

Richard M. Tong
Tarragon Consulting Corporation
1563 Solano Avenue, #350
Berkeley, CA 94707
rtong@tgncorp.com

Ronald R. Yager
Machine Intelligence Institute
Iona College
New Rochelle, NY 10801
yager@panix.com

Abstract

On-line public discussions, such as newsgroups, message boards, and other similar forums, are an under-exploited but potentially valuable resource in developing analyses of world events. An effective way of characterizing this large volume of information is to create time-series that represent the subjects, opinions, and attitudes expressed in these sources. Automatically generated "Linguistic Descriptions" then provide natural and easily understood summaries of the behaviors exhibited by these time series. In this extended abstract, we describe preliminary work on the development of a prototype system to implement our concept of Linguistic Descriptions applied to the attitudinal behaviors seen in on-line discussion forums and news sources.

Operational Motivation

News sources and public opinion polls have long been monitored for indications of public opinion. However, new opportunities exist to understand worldwide reaction to events in the analysis of newsgroups and similar openly available material.

Automated analysis of text, until today, has been primarily focused on extracting and understanding the facts represented in text. Yet attitudinal content, though difficult to gauge, is also central to the intelligence production process.

We believe that large-scale analysis of multiple sources can help to compensate for the biases and particular viewpoints of any individual forum. Large-scale analysis also provides a base-line against which variations can be discovered and evaluated.

The concept of "Linguistic Descriptions" that we introduce in this paper is designed to provide intelligence analysts with a core capability for assessing and interpreting local and global perspectives and their dynamic behaviors.

Linguistic Descriptions

By a Linguistic Description (LD) we mean constructs such as:

> "In the last few weeks, Mid-East sources have become strongly critical of US policy in the region."

and:

> "After a period of recent calm, Indian newspapers have increased calls for action against Pakistani incursions into disputed territory in Kashmir."

We think of these as target constructs to be generated from an analysis of a collection of texts represented as a time series. Formally, we define a Linguistic Description as a 4-tuple:

LD := {T, S, V, B}

where:

Time Extent (T): is the time interval over which the LD holds

Source (S): is the set of sources from which the time series is derived

Variable (V): is the subject or concept that the time series represents

Behavior (B): is the characterization of behavior of the time series over the time extent.

So the first example LD above can be interpreted as:

T: "In the last few weeks"

S: "Mid-East sources"

V: "[criticism of] US policy in the region"

B: "have become strongly [critical]

Note that in this case V is an "attitudinal variable," although the LD model supports variables of any kind. In our core LD model, we define a set of behavior classes that map onto standard behavior descriptors. Current classes include: trends (e.g., up, down), level changes (e.g., higher than, same, lower than), and simple dynamic patterns (e.g., spike, oscillation). Sources are defined with respect to a taxonomy of sources that includes named individuals, as well as named sources and collections of these. Time extents are fuzzy intervals and are typically anchored at one end with respect to a date or an event.

Given this framework, we view the task as one of generating the most useful LD over an underlying database of time-stamped data objects. That is, we interpret the task as a data-mining exercise in which we pre-specify the set of elements from which we can construct the extracted information descriptions.

Attitudes, Opinions, Sentiment

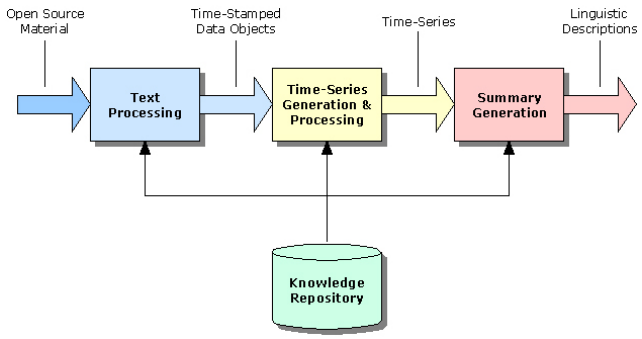
Our approach to the extraction of attitudes and opinions, which we collectively call sentiment, uses a mix of techniques from corpus analysis, machine learning and rule-based pattern matching.

Our experience in the consumer products and services domain suggests that combinations of relatively straightforward techniques can provide us with the level of performance we need to do the large-scale analysis of interest. That is, if the objective is to do "analysis in the large" then we can be less concerned with performance on any specific text and can tolerate higher point error rates than might be considered acceptable in more traditional text classification tasks.

TRENDS™ Infrastructure

Tarragon's TRENDS™ infrastructure is a scalable, configurable content acquisition system that is designed to collect and manage very large amounts of open source material such as that found on websites, message boards, mailing lists, and other Internet forums.

The figure below gives a conceptual overview of the TRENDS™ processing flow.



At each stage of processing we make use of various kinds of knowledge, such as domain ontologies, named entities, and language and data source models.

Specialized modules can be inserted into the processing pipeline as needed. In particular, as improved technologies for attitude and opinion detection become available, these could be straightforwardly incorporated into the TRENDS™ infrastructure.

To illustrate the processing concept, the following text taken from the Islamic Republic News Agency (IRNA), the official news agency of Iran:

```
Tehran, May 23, IRNA -- Iranian Foreign
Minister Kamal Kharrazi here Thursday
described the ongoing situation in the Middle
East "convulsive and critical" and blamed
"repeated US mistakes" for that.
  Talking to IRNA, he said that the US policy
line in the aftermath of the September 11
attack on American landmarks was centered on
meeting the Zionists' interests.
  "The US decision-making and its foreign
policy line is based on meeting illegitimate
goals of the Zionists and this should not be
allowed to lead to instability and collapse
of the international community," Kharrazi
said.
  "The new American policy has adopted use of
pressure as a means to carry other countries
along its side, which has led to the spread
of spite, hatred and war," Kharrazi further
said.
```

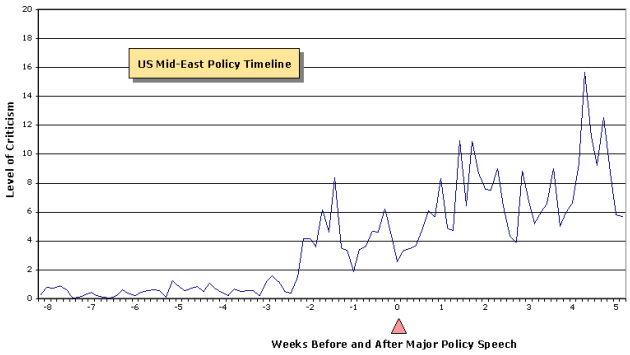
would get mapped into a time-stamped data object that is represented internally using XML, and that would include, in part, a record of our assessment of the sentiment of the text with respect to the issues of interest.

So, in this case we would get something like:

```
<Message id="IAP20020523000110">
<Dtg>23 May 02 1810 GMT</Dtg>
<Source>Tehran IRNA</Source>
<Issue>
  <Issue_Name>US Mid-East Policy</Issue_Name>
  <Issue_Sentiment>-0.80</Issue_Sentiment>
</Issue>
</Message>
```

where the value of -0.80 in the sentiment slot indicates that there is a significant degree of negative sentiment in this message. Sentiment values range over the interval [-1, +1], and, typically, a text would get multiple `<Issue>` tags.

A collection of such data-objects allows us to generate "sentiment timelines" similar to the one shown below:



where we anchor the timeline with respect to some external event (here a major policy speech by the US Foreign Secretary).

This timeline could then give rise to multiple LDs such as:

"Until a few weeks before the speech, very little criticism of US policy was voiced."

"In the weeks immediately preceding the Foreign Secretary's speech, criticism reached new levels."

"Since the speech, criticism has continued to rise."

The intelligence analyst can then interact with the system in a number of ways. For example, they could review the content underlying these summaries, compare results in Mid-East sources with those from, say, Asian sources, and explore changes to the parameterization of the underlying algorithms.

Open R&D Issues

Application of the TRENDS™ concept to the "current events" domain has highlighted a number of open R&D issues that we plan to pursue.

In particular, the empirical evidence we have suggests that, perhaps not surprisingly, detection and quantification of sentiment in this domain is more of a challenge than in the consumer products spaces we have investigated previously (e.g., movies, automobiles, and personal healthcare).

Some of the reasons for this are the highly informal nature of language usage in politically oriented forums coupled with the often chaotic nature of the discourse, and the use of standard language constructs in state-controlled media that need to be calibrated to ascertain their real attitudinal value. In addition, there is the overarching need to assess how effective this analysis is in providing useful intelligence when the external validating events are often obscured or difficult to determine.

Related Work

Our original strategies and techniques for detecting and tracking opinions, as reported by Tong at SIGIR2001, were influenced by the work of Wilson and Rayson on the analysis of transcripts of market research interviews, and by the work of Subasic and Huettner on the development and use of affect lexicons. This work also drew on earlier ideas by Hearst for looking at directionality in text, on the tools for generating lexicons developed by Spertus in her system for recognizing hostile messages, and on the methods used by Sack to characterize discussion themes in Usenet newsgroups.

The study of social networks and on-line communities, and the ways in which they behave and carry ideas, is now a well-established field. Recent books by Rosen, Gladwell and Watts, as well as the edited collection by Smith and Kollock, contain many insights that have contributed to the overall approach adopted in TRENDS™.

More recently we have drawn on the work of Turney and Wiebe to explore more robust attitude and opinion detection techniques, and also on the extensive literature on time-series modeling to create effective time-series segmentation and approximation algorithms.

References

Richard Tong. An Operational System for Detecting and Tracking Opinions in On-Line Discussions. In, *ACM SIGIR 2001 Workshop on Operational Text Classification Systems*, New Orleans, LA, September 2001.

Andrew Wilson and Paul Rayson. The Automatic Content Analysis of Spoken Discourse. In, Clive Souter and Eric Atwell (eds.) *Corpus-Based Computational Linguistics*. Rodopi, 1993.

Pero Subasic and Alison Huettner. Affect Analysis of Text Using Fuzzy Semantic Typing. In *Proc. 9th IEEE International Conference on Fuzzy Systems*. San Antonio, TX, May 2000.

Marti Hearst. Direction-Based Text Interpretation as an Information Access Refinement. In, Paul Jacobs (ed.) *Text Based Intelligent Systems*. Lawrence Erlbaum, 1992.

Ellen Spertus. Smokey: Automatic Recognition of Hostile Messages. In, *Proc. 9th Innovative Applications of Artificial Intelligence*. Providence, RI, July 1997.

Warren Sack. Conversation Map: A Content-Based Usenet Newsgroup Browser. In, *Proc. ACM International Conference on Intelligent User Interfaces*. New Orleans, LA, January 2000.

Emanuel Rosen. *The Anatomy of Buzz*. Doubleday Currency, 2000.

Malcolm Gladwell. *The Tipping Point.* Little, Brown and Co., 2000.

Duncan Watts. *Small Worlds*. Princeton University Press. 1999.

Marc Smith and Peter Kollock (eds.). *Communities in Cyberspace.* Routledge, 1999.

Peter Turney. Thumbs Up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, 2002.

Peter Turney and Michael Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association, *ACM Transactions on Information Systems*, 21 (4), 315-346, 2003.

Jan Wiebe. Learning Subjective Adjectives from Corpora. In, *Proc. 17th National Conference on Artificial Intelligence*, 2000.

Jan Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. Recognizing and Organizing Opinions Expressed in the World Press. In, *AAAI Spring Symposium on New Directions in Question Answering*, Stanford, CA, 2003.