Sentiment analysis

In the novel *Hard Times*, Charles Dickens described the fictional "Coketown" as follows:

> *Fact, fact, fact, everywhere in the material aspect of the town; fact, fact, fact, everywhere in the immaterial. The M'Choakumchild school was all fact, and the school of design was all fact, and the relations between master and man were all fact, and everything was fact between the lying-in hospital and the cemetery, and what you couldn't state in figures, or show to be purchasable in the cheapest market and salable in the dearest, was not, and never should be, world without end, Amen.*

In real life, facts are important, but opinion also plays a crucial role. A computer manufacturer, disappointed with low sales, asks itself: Why aren't consumers buying our laptop? The Democratic National Committee, disappointed with the last election, wants to know on an on-going basis: What is the reaction in the press, newsgroups, chat rooms, and blogs to Bush's latest policy decision?

Answering these questions requires focusing on subjective judgments (e.g. the design is tacky, the administration ignored previous treaties) while taking into account misperceptions (e.g. updated device drivers aren't available), the effect of indirect reporting (e.g. Bush assured the crowd that European support was broad), and the existence of possibly conflicting opinions from the same person or organization.

CS professors Claire Cardie and Lillian Lee are working on sentiment-analysis technologies for extracting and summarizing *opinions* from unstructured human-authored documents. They envision systems that (a) find reviews, editorials, and other expressions of opinion on the Web and (b) create condensed versions of the material or graphical summaries of the overall consensus.

CS Professors Claire Cardie (left) and Lillian Lee (above): Be cautious when you hear, "It is a fact that … ."; the phrase is highly correlated with the introduction of an opinion rather than a fact!

Indeed, the Cornell Natural Language Processing group has done seminal work in developing algorithms for sentiment classification and extraction problems, and its research has been widely recognized in the research community and in the scientific popular press as being, in large part, responsible for the recent huge surge of interest in the area.

Over a dozen external groups have written papers using the so-called Cornell movie-review dataset as a benchmark.

Problems considered by the group include the following: determine whether a document or portion thereof is subjective, determine whether the opinion expressed is positive or negative, determine the strength of the sentiment (e.g. is France *really* or just mildly unhappy with Bush?), find the sources of the opinion (the person, group, report, etc.), and determine whether the opinion is being filtered through indirect sources (e.g. as "Bush" took the liberty of attributing an opinion to "Europeans" in the example above). At first glance, this might not appear so hard. For example, can't one just look for obvious *sentiment indicators* —words like "great"?

The difficulty lies in the richness of human language use. The amazingly large number of ways to say the same thing (especially, it seems, when that thing is a negative perception) complicates the task of finding a high-coverage set of indicators. Furthermore, the same indicator may admit several different interpretations, as the following sentences show:

- This laptop is a great deal.
- A great deal of media attention surrounded the release of the new laptop model.
- If you think this laptop is a great deal, I've got a nice bridge for you to buy.

Each of these sentences contains the phrase "a great deal", but the opinions expressed are, respectively, positive, neutral, and negative. The first two sentences use the same phrase to mean different things. The last sentence involves sarcasm, which, along with related rhetorical devices, is an intrinsic feature of texts on blogs, newsgroup postings, and, more generally, opinionated text.

Researchers have adopted basically two approaches to meeting the challenges of sentiment analysis. Many

groups are working to incorporate linguistic knowledge; given the subtleties of natural language, such efforts will be critical to building operational systems.

Cardie and Lee pursue a different tack: they employ learning algorithms that can automatically infer from text samples what word-level indicators and phrase-based syntactic and semantic patterns are useful for sentiment analysis.

Learning systems are potentially more cost-effective, more easily ported to other domains and languages, and more robust to grammatical mistakes. Furthermore, they can discover indicators and patterns that humans might neglect. Lee's group found, for example, that, in certain types of text, the phrase "still," (comma included) is a better indicator of positive sentiment than "good" —a typical use is, "Still, despite these flaws, I'd go with this laptop." And, Cardie and her collaborators at the Universities of Utah and Pittsburgh found that the pattern "It is a fact that …" is highly correlated with the introduction of an opinion rather than a fact!

Given the multitude of potential applications, researchers like Cardie and Lee have been devoting more and more attention to sentiment analysis. If they continue to be successful, their systems could save information analysts from having to read and summarize potentially hundreds of documents for each topic of interest and would save analysts at the aforementioned laptop manufacturer from having to read potentially hundreds of versions of the same complaints. Surely that sounds like a great deal!

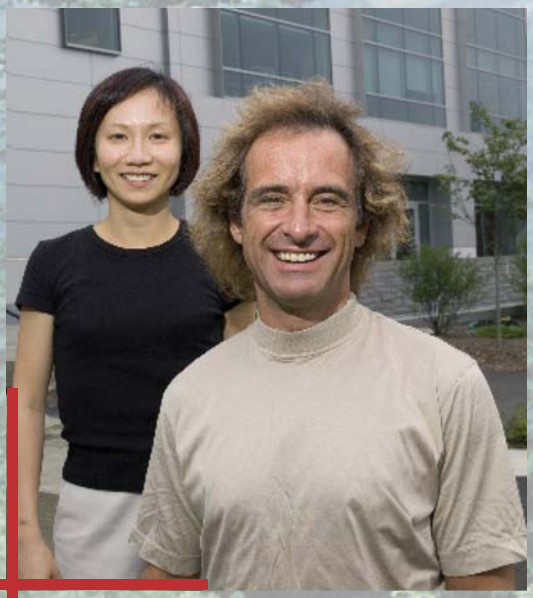
CURIE comes to CS

The CURIE Academy is a one-week Cornell residential program for high school girls who excel in math and science. The students work in teams on a carefully formulated project designed to develop their problem-solving skills and immerse them in an interdisciplinary topic. There is, of course, time for introductions to many other areas of Engineering at Cornell and for non-academic fun.

Over the past eight years, the CURIE program has drawn a strong and diverse applicant pool from across the country. Most of the students go on to top-ranked colleges for science and engineering. Many past CURIE members who come to study at Cornell act as undergrad facilitators and mentors to the new crop.

CS faculty Graeme Bailey and Daisy Fan are on the board that oversees CURIE. In 2000, Fan headed up a highly successful CURIE project on environmental systems modeling, with challenging engineering and ethical problems to engage them. In summer 2005, the focus was on computer graphics, and around 40 high school girls did a project with CS faculty Kavita Bala and Steve Marschner.

In 1865, Ezra Cornell said, "I would found an institution where any person can find instruction in any study." More girls than ever are studying engineering and science high above Cayuga's waters, thanks to the CURIE program.



I had a wonderful experience at the CURIE Academy. [It] enabled me to spend a week with the most incredible girls who share my passion for math and science.

~ Isabelle Puckette
16, Encinitas, CA

Daisy Fan (left) and Graeme Bailey are on the Board of Directors of the CURIE program.

Srinivas Keshav, Greg Morrisett, Praveen Seshadri, David Shmoys join.

1996

Don Greenberg receives the ASCA Creative Research Award in Architecture.

Dan Huttenlocher receives a Cornell Presidential Weiss Fellowship for his contributions to undergraduate education. Three such awards are given each year; Cornell has 1600 faculty members.

David Gries receives an honorary doctorate from Daniel Webster College in New Hampshire.

Bruce Land gets first place in the instructional materials (Web-based) competition of the ACM SIGUCCS Use Services Conference XXIV. The award was for the Web site for his graphics programming course: http://instruct1.cit.cornell.edu/courses/cs418-land.

Joe Halpern becomes Editor-in-Chief of the *Journal of the ACM*.

Graeme Bailey, Lillian Lee, Bart Selman join. CS grows to 30 faculty and has over 500 computers.

1997

Juris Hartmanis takes a two-year leave to serve as Assistant Director of the NSF for CISE. During his tenure, he effectively positions NSF and CISE to assume a leadership role in response to the PITAC report, and he is instrumental in shaping the discussion that lead to NSF's playing the lead role in the Information Technology Research (ITR) program.

Joe Halpern shares the 1997 Gödel Prize with former student Yoram Moses. Their paper *Knowledge and Common Knowledge in a Distributed Environment*, says the citation, "provided a new and effective way of reasoning about distributed systems".

David Shmoys becomes Editor-in-Chief of the SIAM *Journal of Discrete Mathematics*.

The faculty publish six books:
Ken Birman, *Building Secure and Reliable Network Applications* (Prentice Hall).

Srinivas Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network* (Addison-Wesley).

Dexter Kozen, *Automata and Computability* (Springer-Verlag).

Fred Schneider, *On Concurrent Programming* (Springer-Verlag).

Nick Trefethen and student David Bau, *Numerical Linear Algebra* (SIAM).