FAS

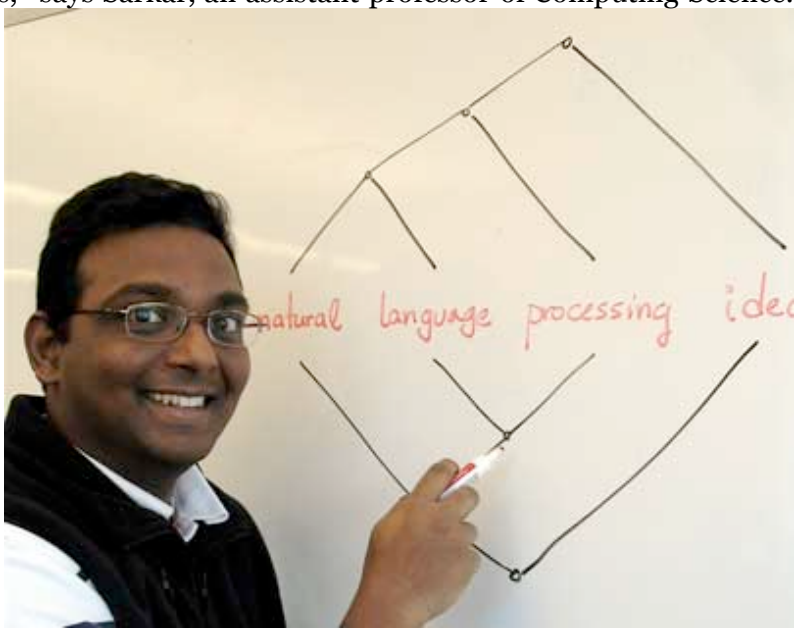
40 YEARS FACULTY OF APPLIED SCIENCES SIMON FRASER UNIVERSITY

Understanding Language Understanding

A man bought a shirt with pockets. He bought the shirt with a credit card. What makes you assume the shirt has pockets, but not that the fellow handed over a pile of pockets for a shirt with a credit card sewn into it? Computer scientists must tackle such problems before they can teach computers to understand language.

Anoop Sarkar [http://www.cs.sfu.ca/%7Eanoop/] finds it amusing that he and his graduate students are developing software that could write this article some day. The project called SQuASH (SFU Question Answering Summary Handler) scans multiple documents such as newspapers or academic abstracts and then, based on a set of questions provided by the user, creates a short readable summary.

According to him, journalists should start worrying. "Our natural language processing ideas are modelled after the way babies learn language. They learn rapidly from input they observe because they combine important prior knowledge with novel experiences," says Sarkar, an assistant professor of Computing Science.



As a teenager Sarkar was fascinated by computer programming languages. He says, "They are designed to be simple and easy to process so that each program has exactly one meaning. But they are artificial languages." In contrast, natural human languages are complex and full of ambiguity. At Pune University in India Sarkar wrote a compiler [http://en.wikipedia.org/wiki/Compilers] which translated computer programs into low level instructions for computers. "I felt inspired to try to find similar tools for natural languages," says Sarkar. That led him to graduate school at the University of Pennsylvania with professor Aravind Joshi, a pioneer in the field. Joshi wrote one of the first programs to analyze natural language on a UNIVAC-1 (it had 1K of memory!).

Take the phrase, "natural language processing ideas". It has five possible structures with different meanings. In the picture, at the whiteboard, he demonstrates ambiguity in language using trees. The tree on top is the more plausible one. Here are all the possible structures:

((natural language) (processing ideas)) ==> various kinds of processing ideas applied to natural language
(natural (language (processing ideas))) ==> a natural set of processing ideas applied to language
(((natural language) processing) ideas) ==> ideas about processing natural language
(natural ((language processing) ideas)) ==> a natural set of ideas about language processing
((natural (language processing)) ideas) ==> a set of ideas about natural methods in language processing

Adding one more word (e.g. natural language processing research ideas) yields 14 possible meanings. Nine words combine in 1,430 ways. How do you pick the right one? Most humans would naturally pick the third meaning above. But why?

Sarkar's research helps computers learn how to choose. "We expose our learning software to hundreds of thousands of cases for which the most plausible meaning is provided by humans," says Sarkar. But this is not enough. For effective machine learning the software is shown additional examples, analogous to the way a baby observes language without supervision. The computer has statistical prior knowledge to make educated guesses that a given meaning is probably the most plausible one.

This core set of statistical 'natural language processing ideas' can be applied to translation, mining information from text, and summarization. So, if a future computer could write an article like this one, would you still read it? Or will you just ask your computer to read it for you and give you a summary?

Example output from SQuASH is viewable on Sarkars website [http://www.cs.sfu.ca/%7Eanoop/distrib/merged_2.4.xml] . NOTE: the summaries are generated from 25 - 50 preselected documents provided by the US National Institute of Standards and Technology (NIST) from two sources: The Financial Times of London, and The Los Angeles Times. Pre-selection is done semi-automatically by NIST, using information retrieval methods and a human assessor. The goal is to produce a readable short summary from a set of relevant documents, as opposed to information retrieval which finds relevant documents from a vast set of documents that may or may not be relevant to a query. The data is comes from an annual competition on summarization held by NIST [http://duc.nist.gov/duc2005/] . for more information.