Computer Translator Reads Between The Tweets

by CHRISTOPHER JOYCE

April 4, 2011

One way to follow what's going on in the Middle East and South Asia right now is through social media — Facebook, Twitter and blog posts.

But of course you have to speak the local languages to do that. So scientists in the U.S. are trying to get computers to work around that problem.

There are already computer programs to translate text from one language to another. But languages like Arabic or Urdu are tough — the script and the grammar are hugely different from European languages, and digitized dictionaries and grammatical algorithms for those languages are still in the early stages.

But these languages are politically important. Urdu, for example, is the language of Pakistan and of many Muslims in India. It's a mix of Hindi and Persian and uses Arabic script.

When you are able to figure out what the topic of the conversation is, what kind of sentiment is being expressed around that, that's the goal of what we are trying to do.

- Rohini Srihari, computer scientist, University of Buffalo

Mining For Sentiment And Opinion

Computer scientist Rohini Srihari says existing computer translators for Urdu are often too literal.

"What I want is to determine who are the people, places and things being talked about," she says. "Is there an opinion being expressed? Is it a positive or negative opinion being expressed?"

At the University of Buffalo, Srihari has developed a natural language program that she says can do that. The computer has "learned" the nuances of written Urdu. Some of this is fairly mechanical — Urdu doesn't have the same kind of clear breaks between one word and the next the way English does. Some things are subtle, as in characters and words whose placement may connote sentiment.

"And when you are able to figure out what the topic of the conversation is," she says, "what kind of sentiment is being expressed around that, that's the goal of what we are trying to do."

On the screen, you can mouse over a section of script and if it carries a "negative" connotation — it will highlight red. If it has a "positive" sentiment, it glows green.

Related NPR Stories

Twitter's Biz Stone On Starting A Revolution

The co-founder of Twitter discusses how his platform has been used to spread information worldwide.

Twitter Turns Five: #happybirthday!
March 19, 2011

The Revolution Will Be Tweeted
Feb. 21, 2011

A Digital Revolution In The Palm Of Your Hand

July 21, 2010

Srihari says the computer allows her to mine the Internet. Her research company, called Janya, Inc., gets funding from the Pentagon for the project.

"So in Twitter posts and tweets and so on, if there's specific factual information that's being mentioned — they want that extracted," says Srihari. "There's also definitely an interest in sentiment and opinion mining."

Getting More Voices Into The Conversation

Social media and the opinions that get slung around digitally around the world's hot spots are also of interest to political scientists and historians.

Ernest Tucker, a history professor with the Center for Middle East and Islamic Studies at the U.S Naval Academy, struggles with Urdu himself. But Tucker does speak and read Persian, which is close, and he regularly reads publications from the region.

He argues that history is best told not by what the Napoleans say, but by the foot-soldiers, or, in this case, the tweeters.

"And that's the goal of all historians anywhere," he says, "to try to get the voices of more and more people into the conversation, and anything that can do that, particularly this kind of thing, is a wonderful gift."

That's the goal of all historians anywhere — to try to get the voices of more and more people into the conversation.

- Ernest Tucker, history professor, The Center for Middle East and Islamic Studies, U.S Naval Academy

Tucker says he's skeptical about how well a computer is going to identify sentiment — he says you'll still need a human linguist to fine-tune any translation. For example, he says it's common in Middle Eastern languages to employ couplets from traditional poetry to convey feelings — symbolic language that could confuse a computer program.

"For the Iranians, for the Pakistanis, for the Indians," he says, "it's still part of the living connection to the cultures of the past."

Rohini Srihari acknowledges the program isn't perfect. It gets flummoxed by things like Urdish, a mashup language for text messaging that's part Urdu, and part English. But it has given her insight into what Urdu speakers have been talking about lately.

"A lot of the conversation, believe it or not, was about cricket — that seems to be on everyone's mind al the time," she says.

Last week would have been a good time to tune into Urdu cyberspace. Pakistan played India in the world cricket semi-finals. Pakistan lost; no doubt the web traffic was full of strong sentiments.