Mining the blogosphere—Researchers develop tools that make sense of social media

September 6 2012

Can a computer "read" an online blog and understand it? Several Concordia computer scientists are helping to get closer to that goal.

Leila Kosseim, associate professor in Concordia's Faculty of Engineering and Computer Science, and a recently-graduated doctoral student, Shamima Mithun, have developed a system called BlogSum that has potentially vast applications. It allows an organization to pose a question and then find out how a large number of people talking online would respond. The system is capable of gauging things like consumer preferences and voter intentions by sorting through websites, examining real-life self-expression and conversation, and producing summaries that focus exclusively on the original question.

"Huge quantities of electronic texts have become easily available on the Internet, but people can be overwhelmed, and they need help to find the real content hiding in the mass of information," explains Kosseim, one of the lead researchers at Concordia's Computational Linguistics Laboratory (CLaC lab).

Analyzing informally-written language poses unique challenges compared to analyzing, for example, a news article. Blogs, forums and the like contain opinions, emotions and speculations, not to mention spelling errors and poor grammar. A summarization tool must address two particular problems, question irrelevance (sentences that are not relevant to the main question), and discourse incoherence, (sentences in which the intent of the writer is unclear).

BlogSum met these challenges with demonstrable efficiency. The researchers developed and tested their tool by examining a set of blogs and review sites. BlogSum used "discourse relations" to crunch the data – ways of filtering and ordering sentences into coherent summaries. BlogSum was measured against prior computational rankings and achieved mostly superior results. In addition, it was evaluated by actual human subjects, who also found it to be superior. Summaries produced by BlogSum reduced question irrelevance and discourse incoherence, successfully distilling large amounts of text into highly readable summaries.

This study is an example of Natural Language Processing (NLP), in which Concordia, through the CLaC lab, is a leader. NLP stands at the intersection of artificial intelligence and linguistics, seeking to enable computers to derive meaning from human language.

"The field of natural language processing is starting to become fundamental to computer science, with many everyday applications – making search engines find more relevant documents or making smart phones even smarter," explained Kosseim.

Provided by Concordia University