Researchers design machine learning technique to improve consumer medical searches

November 17 2010

Medical websites provide consumers with more access than ever before to comprehensive health and medical information, but the sites' utility becomes limited if users use unclear or unorthodox language to describe conditions in a site search. However, a group of Georgia Tech researchers have created a machine-learning model that enables the sites to "learn" dialect and other medical vernacular, thereby improving their performance for users who use such language themselves.

Called "dia™" (short for "dialect topic modeling"), the system learns by comparing multiple medical documents written in different levels of technical language. By comparing enough of these documents, dia eventually learns which medical conditions, symptoms and procedures are associated with certain dialectal words or phrases, thus shrinking the "language gap" between consumers with health questions and the medical databases they turn to for answers.

"The language gap problem seems to be the most acute in the medical domain," said Hongyuan Zha, professor in the School of Computational Science & Engineering and a paper co-author. "Providing a solution for this domain will have a high impact on maintaining and improving people's health."

To educate dia in various modes of medical language, Crain and his fellow researchers pulled publicly available documents not only from

WebMD but also Yahoo! Answers, PubMed Central, the Centers for Disease Control & Prevention website, and other sources. After processing enough documents, he said, dia can learn that the word "gunk," for example, is often a vernacular term for "discharge," and it can process user searches that incorporate the word "gunk" appropriately.

In this initial study using small-scale experiments, the researchers found that dia can achieve a 25 percent improvement in nDCG ("normalized discounted cumulative gain"), a scientific term that refers to the relevance of information retrieval in a web search. Zha, whose research focuses on Internet search engines and their related algorithms, said a 5 percent improvement in nDCG is "very significant."

"Dia figures out enough language relationships that over time it does quite well," said Steven Crain, Ph.D. student in computer science and lead author of the paper that describes dia. "Another benefit is we're not doing word-for-word equivalencies, so 'gunk' doesn't necessarily have to be connected to 'discharge,' as long as it's recognized that 'gunk' is related to infections."

Also, dia is not limited to medical search; it is a machine-learning technique that would work equally well in any topic-related search. In addition to approaching websites about incorporating dia into their search engines, Crain said one next stop is to develop the model so that it can learn dialects by looking at patterns that do not make sense from a topical perspective. For example, using a similar algorithm he was able to automatically discover dialects including text-speak dialect (e.g. "b4" as a subsititue for "before"), but the dialects were mixed in with topically-related groups of words.

"We're trying to get to where you can isolate just the dialects," Crain said.

"This feature will help common users of medical websites," Zha said. "It will help enable consumers with a relatively low level of health literacy to access the critical medical information they need."

Provided by Georgia Institute of Technology