Detecting Twitter users' gender, en francais

November 28 2013, by Chris Chipello

With 230 million users, Twitter has become a global force in social media. And not just in English.

Data miners have been hard at work trying to figure out the attributes of Twitter users – such as gender and age—that aren't explicitly revealed on Twitter feeds. That information could be hugely valuable to marketers, enabling them to target messages to their desired audience. Nearly all the research done so far, however, has focused on English users and content.

Now, a McGill University research team has conducted one of the first studies designed to figure out the gender of Twitter users who primarily use languages other than English.

Among the key findings: by using a special detector based on French-language syntax, the researchers showed that it is very easy to classify gender for Twitter users in French – and probably for other Romance languages. In particular, the researchers developed an algorithm to look for masculine or feminine adjectives or past participles following the phrase "Je suis" (or variants such as "je ne suis pas").

Based on this construction, the detector was able to determine the gender of users with 90% accuracy – significantly higher than the accuracy rates of 80% to 85% achieved by various algorithms that have been developed to analyze English-language content.

Because French adjectives and past participles have masculine and feminine forms that are often spelled differently, "You don't have to get

too fancy" to develop an effective gender detector for Tweets in the language, says Derek Ruths, a McGill computer-science professor who co-authored the study.

Since most individuals include photos of themselves on their Tweets, identifying male and female users might seem as simple as looking at the photos. But sorting through hundreds of millions of tweets is a task for computers, and "computers aren't good at looking at pictures," Ruths notes.

The McGill study was presented at a recent international conference in Seattle organized by the Association for Computational Linguistics. The paper also examines Twitter data sets for Japanese, Indonesian and Turkish. Japanese proved to be the toughest for inferring gender.

The results obtained for French show that some languages have features better suited for certain classification tasks. "Identifying and leveraging such features promises to be an interesting and effective direction for future work," adds McGill linguistics professor Morgan Sonderegger, who co-authored the paper with Ruths and computer-science undergraduate student Morgane Ciot.

More information: Link to the paper: www.derekruths.com/static/publ … rRuths_EMNLP2013.pdf
Link to the conference website: hum.csse.unimelb.edu.au/emnlp2013/

Provided by McGill University