Introducing diversity in online language analysis

September 7 2016, by Janet Lathrop

Brendan O'Connor is an expert in natural language processing at UMass Amherst. Credit: UMass Amherst

For the past 30 years, computer science researchers have been teaching their machines to read, for example, assigning back issues of the Wall Street Journal, so computers can learn the English they need to run search engines like Google or mine platforms like Facebook and Twitter for opinions and marketing data.

But using only standard English has left out whole segments of society who use dialects and non-standard varieties of English, and the omission is increasingly problematic, say researchers Brendan O'Connor, an expert in natural language processing (NLP) at the University of Massachusetts Amherst, and Lisa Green, director of the campus' Center for Study of African-American Language. They recently collaborated with computer science doctoral student Su Lin Blodgett on a case study of dialect in online Twitter conversations among African Americans.

Details appear in their paper posted online now in advance of their presentation at the Empirical Methods on NLP conference on Nov. 2-5 in Austin, Texas. The authors believe their study has created the largest data set to date for studying African-American English from online communication, examining 59 million tweets from 2.8 million users.

As O'Connor explains, "We have a huge amount of digital information now that we didn't have before, and many different demographic groups are now using new technologies. On the computer science engineering side, a lot more types of people are using search engines like Google, and the computer needs to be able to parse the text to understand what they're asking."

On the social side, Green adds, people from many different social groups use different language than is found in mainstream media, especially casually or among themselves. She notes, "New semantics can be expanded very quickly if some expression is picked up from dialect by the larger community. As linguists, we are always interested in how language changes and now we are seeing some changes happening very quickly. For example, consider the expression 'stay woke' on Twitter."

O'Connor says, "What's interesting now is that all this important textual data is being generated in a less formal context. If we want to analyze opinions about an election, for example, we still use NLP tools to do it, but right now, the tools are all geared for standard, formal English. There are clearly deficiencies in status quo technologies."

To expand NLP and teach computers to recognize words, phrases and language patterns associated with African-American English, the researchers analyzed dialects found on Twitter used by African Americans. They identified these users with U.S. census data and Twitter's geo-location features to correlate to African-American neighborhoods through a statistical model that assumes a soft correlation between demographics and language.

They validated the model by checking it against knowledge from previous linguistics research, showing that it can successfully figure out patterns of African-American English. Green, a linguist who is an expert in the syntax and language of African-American English, has studied a community in southwest Louisiana for decades. She says there are clear patterns in sound and syntax, how sentences are put together, that characterize this dialect, which is a variety spoken by some, not all, African Americans. It has interesting differences compared to standard American English; for example, "they be in the store" can mean "they are often in the store."

The researchers also identified "new phenomena that are not well known in the literature, such as abbreviations and acronyms used on Twitter, particularly those used by African-American speakers," notes Green. adds, "This is an example of the power of large-scale online data. The size of our data set lets us characterize the breadth and depth of language."

Finally, the researchers evaluated their model against existing language classifiers to determine how well existing NLP tools perform in analyzing African-American English in user-level and message-level analyses. They found that current widely used tools identify African-American English as "not English" at higher rates than expected, O'Connor says. Testing the best open source language classification software and Twitter's own language identifier, they found the open source system was almost twice as bad for African-American English than for online English associated with whites in the U.S. The researchers also found similar issues with Google's state-of-the-art SyntaxNet grammatical parser.

He adds, "These methods are used by Google and other companies on millions of web pages every day to extract meaning for systems like search engines. Since African-American English is analyzed poorly, that implies information access is worse for texts authored by African-American English speakers. The issue of fairness and equity in artificial intelligence methods is of increasing concern, since they are crucial to technologies we use every day, like search engines."

Furthermore, O'Connor states, "Technology companies have well-known issues with diversity. For example, Facebook and Google recently reported that only 2 percent of their employees are African-American. Hopefully, efforts to increase diversity among technologists can help draw attention to addressing problems of fairness in artificial intelligence."

For her part, Green hopes the new model will show that "there might be new opportunities for young African-American English speakers to contribute further to natural language processing. We might be able to look forward to attracting more African-American English speakers, and members of other underrepresented groups, to engineering and computer science." The authors plan to release their new model in the next year to better identify English written in these dialects by using publicly available data from Twitter.

More information: Demographic Dialectal Variation in Social Media: A Case Study of African-American English, arXiv:1608.08868 [cs.CL] arxiv.org/abs/1608.08868

Provided by University of Massachusetts Amherst