LEXICON VALLEY A BLOG ABOUT LANGUAGE

MARCH 16 2015 11:23 AM

Artificial Belligerence

By Iva Cheung

*This **post** originally appeared on **Strong Language**, a sweary blog about swearing.*

What makes us human? Our innate curiosity? Our mastery of language? Or is it our astounding ability to be complete assholes to one another?

As an undergraduate student in the late 1980s, computer scientist **Mark Humphrys** programmed an **ELIZA**-like chatterbot and discovered that it could pass as human when he gave it a mean streak:

> The original Eliza was meant to be sympathetic to the human. I thought it would be interesting to add aggression and profanity to the program. My program was deliberately designed to have an unpredictable (and slightly scary) mood.

His bot would respond to innocuous statements or questions with such abusive remarks as "you are obviously an asshole" or "ok honestly when was the last time you got laid?" If the person became angry and began flinging insults, the bot would come back with "you only use foul language to make up for your small penis."

"Here you go, idiot."

Photo by STR/AFP/Getty Images

Humphrys documents one particular conversation, with a user called SOMEONE at DRAKE, which shows the boorish bot having passed the **Turing test**, "depending on how one defines the rules (and how seriously one takes the idea of the test itself)."

The key, explains Humphrys, was to push the right emotional buttons to cloud the user's judgment: "Questioning SOMEONE at DRAKE's sexual abilities was all that was needed to get him into defensive, boastful and argumentative mood, emotions that blinded him to the fact that he wasn't actually talking to a person."

Another means to the same end is to use sex directly, and Humphrys, with deep admiration, tells about Jake Kaufman's Jenny18 bot, which, posing as a horny 18-year-old woman, brings user dom01 "to orgasm repeatedly, which surely must count as a fundamental new milestone passed in the history of the Turing Test."

These very specific instances aside, slang and swearing have proven tricky for artificial intelligence and natural language processing researchers to implement on a machine. Although curse words make up only **0.5 to 0.7 percent** of all of the words we speak, **they are rich in nuance and play a variety of roles**. Said IBM research scientist Eric Brown, "As humans, we don't realize just how ambiguous our communication is."

In 2011 Brown's team tried to train *Jeopardy!*-winning supercomputer Watson to **use more natural-sounding vocabulary** by feeding it the entirety of **Urban Dictionary**. The result was a foul-mouthed machine that learned terms such as "ass hat" and "front butt" but didn't understand when it was appropriate to use them, once responding to a researcher's query with "bullshit!" Watson's failure to distinguish between profanity and polite language meant that Brown's team had to develop filters to screen out the profanity and eventually ended up wiping Urban Dictionary's entries from Watson's memory.

Working on the receiving end of belligerent speech acts is an **international team of researchers** hoping to develop bots that can take—or, rather, *not* take—the worst you can dish out. Applying politeness theory, a branch of linguistic pragmatics, and its offshoot, *impoliteness* theory (that is, not merely failing to be polite but succeeding at being rude), Ronald M. Lee and Gregory Diaz of Florida International University, along with Elizabeth Dominguez Campillo of the Universidad de la Habana in Cuba, have proposed artificially intelligent bots that could become offended by a variety of **face-threatening acts**—not only deliberate insults but also inadvertent lapses in tact. Because each culture has different rules for what's considered polite, we'd theoretically be able to feed a bot the rules for a specific culture and use it as a training tool for business executives or politicians on diplomatic missions. Like humans, these bots would take offense and escalate in their anger if you didn't treat them properly. "However," said the researchers, "they have one important redeeming grace: when asked, they will explain to you why they are angry."

Machines that can be trained to recognize aggression, insults, and swearing have other applications—we might want to censor profanity in certain situations, for instance, or flag instances of cyberbullying. But existing systems, based on lists of potentially offensive words, perform poorly: According to California-based computer scientists **Sara Owsley Sood and Judd Antin**, “They fail to adapt to evolving profane slang, identify profane terms that have been disguised or only partially censored (e.g., @ss, f$#%) or intentionally or unintentionally misspelled (e.g., biatch, shiiiit).* For these reasons, they are easy to circumvent and have very poor recall.” Further, words such as *ass, cock*, and ***cum*** can be indecorous in some situations but not in others—context is essential. Sood and Antin used a crowdsourced data set of profanity to encourage **support vector machines**, used in text classification, “to learn a model of profanity.” The researchers’ profanity detection system performed much better than a traditional list-based system, but they admit they still have work to do to improve recall.

When it comes to language, especially the pragmatics of strong language, artificial intelligence hasn’t caught up with humans yet. But with so many A.I. researchers attacking the problem from different angles, a future with realistic potty-mouthed sass-bots that can verbally spar with the best (worst?) of us may not be that far away.

***Correction, March 16, 2015**: *This post originally misspelled the last name of researcher Sara Owsley Sood.*

NEWS & POLITICS

JURISPRUDENCE