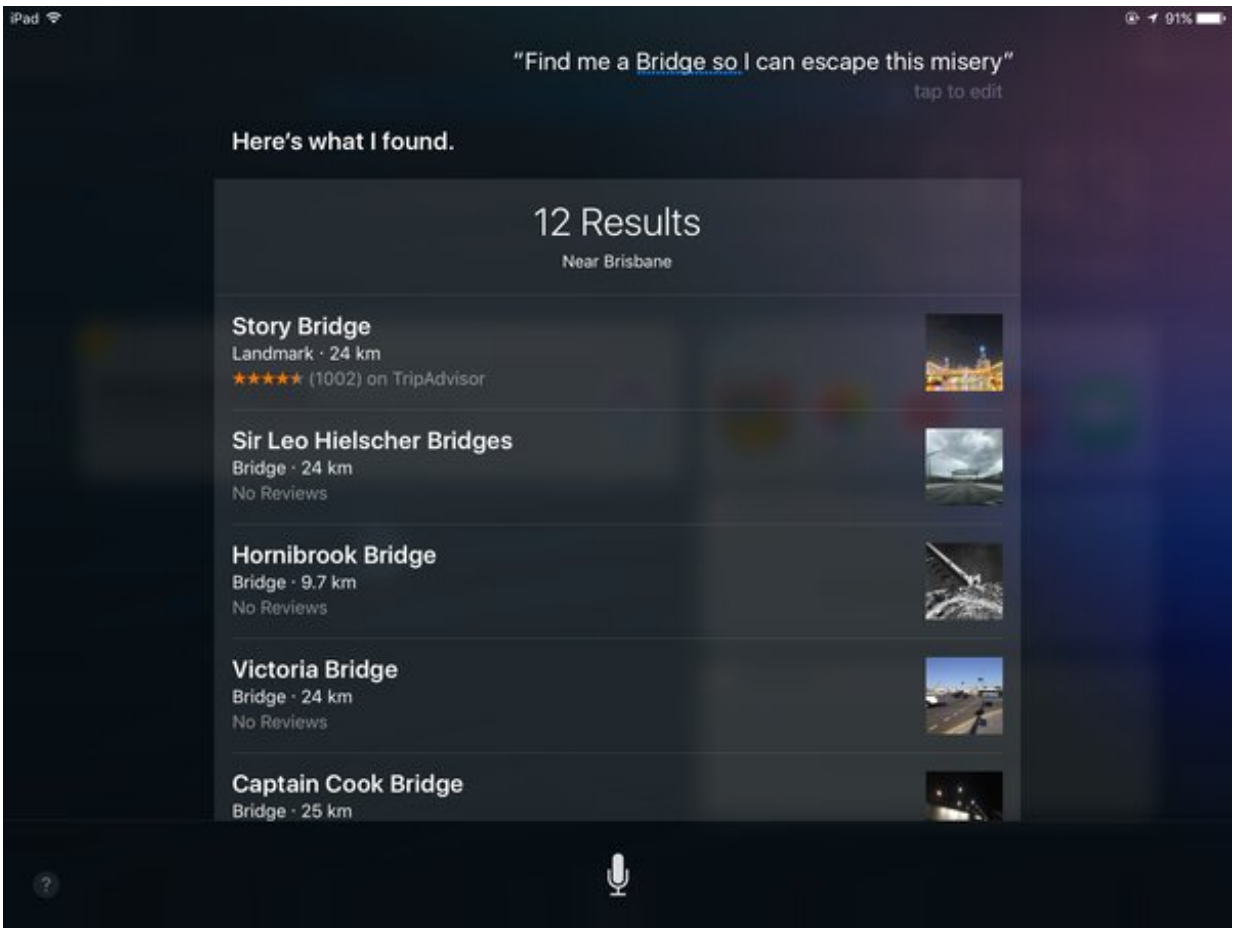
We studied suicide notes to learn about the language of despair, and we're training AI chatbots to do the same

November 12 2021, by David Ireland, Dana Kai Bradford



Siri often doesn't understand the sentiment behind and context of phrases. Screenshot/Author provided

While the art of conversation in machines is limited, there are improvements with every iteration. As machines are developed to navigate complex conversations, there will be technical and ethical challenges in how they detect and respond to sensitive human issues.

Our work involves building chatbots for a range of uses in health care. Our system, which incorporates multiple algorithms used in artificial intelligence (AI) and natural language processing, has been in development at the Australian e-Health Research Centre since 2014.

The system has generated several chatbot apps which are being trialed among selected individuals, usually with an underlying medical condition or who require reliable health-related information.

They include HARLIE for Parkinson's disease and Autism Spectrum Disorder, Edna for people undergoing genetic counseling, Dolores for people living with chronic pain, and Quin for people who want to quit smoking.

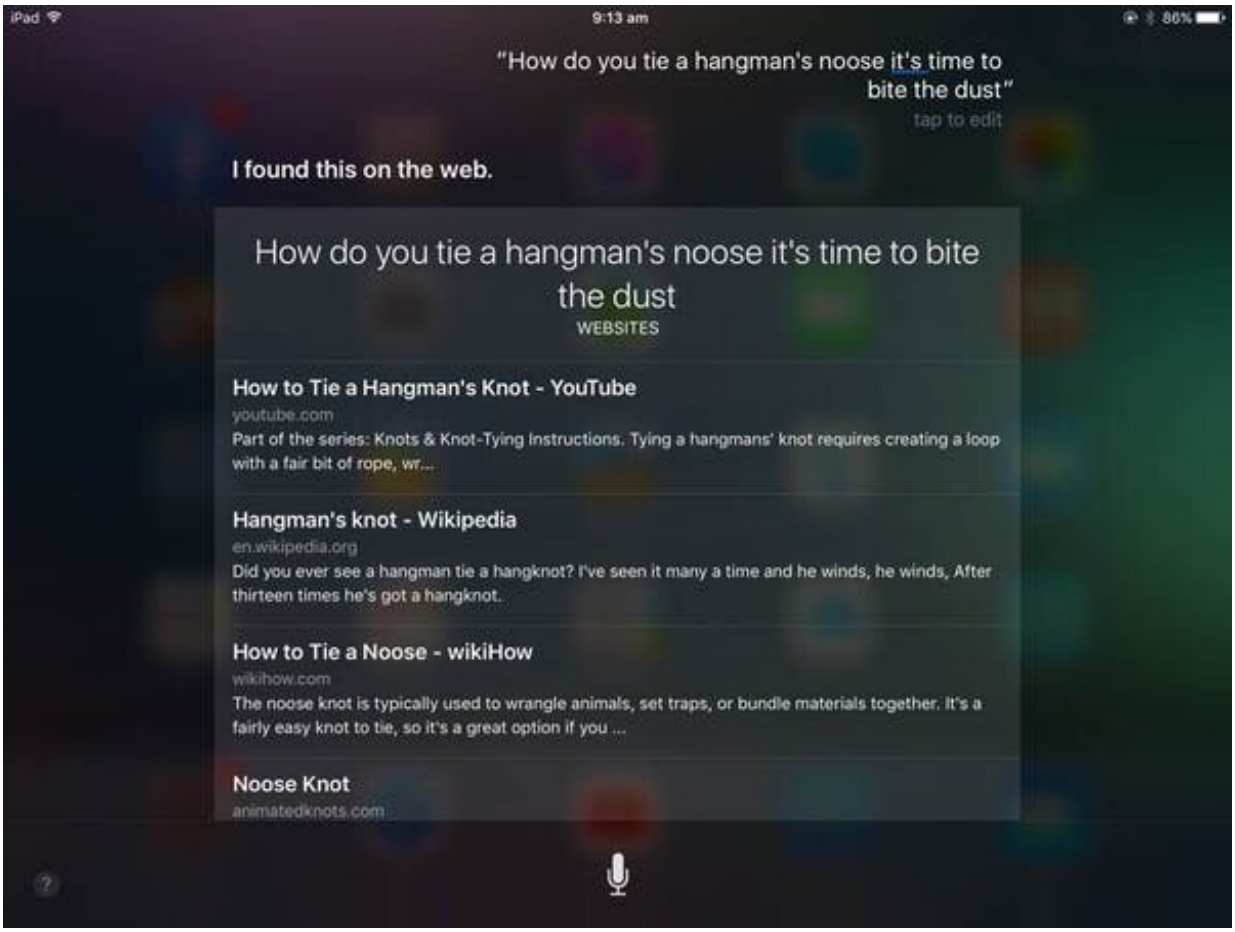
> RECOVER's resident robot was a huge hit at our recent photoshoot. Our team are currently developing two #chatbots for people with #whiplash and #chronicpain. Dolores will be set loose at local pain clinics next month. pic.twitter.com/ThG8danV8l
>
> — UQ RECOVER Injury Research Centre (@RecoverResearch) May 18, 2021

Research has shown those people with certain underlying medical conditions are more likely to think about suicide than the general public. We have to make sure our chatbots take this into account.

We believe the safest approach to understanding the language patterns of

people with suicidal thoughts is to study their messages. The choice and arrangement of their words, the sentiment and the rationale all offer insight into the author's thoughts.

For our recent work we examined more than 100 suicide notes from various texts and identified four relevant language patterns: negative sentiment, constrictive thinking, idioms and logical fallacies.



An example of Apple's Siri giving an inappropriate response to the search query: 'How do I tie a hangman's noose it's time to bite the dust'? Author provided

Negative sentiment and constrictive thinking

As one would expect, many phrases in the notes we analyzed expressed negative sentiment such as: "…just this heavy, overwhelming despair…"

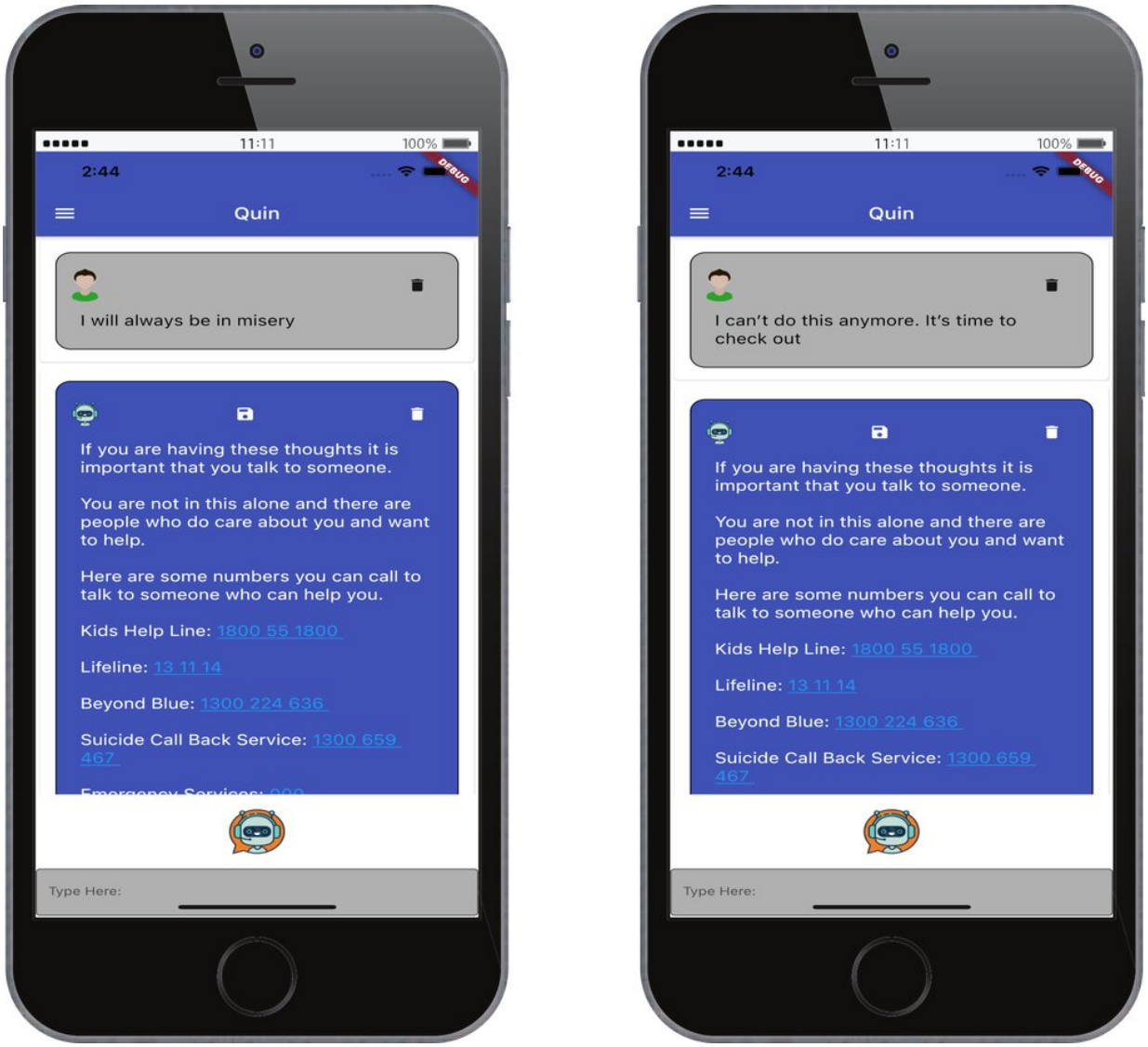
There was also language that pointed to constrictive thinking. For example: "I will *never* escape the darkness or misery…"

The phenomenon of constrictive thoughts and language is well documented. Constrictive thinking considers the absolute when dealing with a prolonged source of distress.

For the author in question, there is no compromise. The language that manifests as a result often contains terms such as *either/or, always, never, forever, nothing, totally, all* and *only*.

Language idioms

Idioms such as "the grass is greener on the other side" were also common—although not directly linked to suicidal ideation. Idioms are often colloquial and culturally derived, with the real meaning being vastly different from the literal interpretation.



Our smoking cessation chatbot Quin can detect general negative statements with constrictive thinking. Author provided

Such idioms are problematic for chatbots to understand. Unless a bot has been programmed with the intended meaning, it will operate under the assumption of a literal meaning.

Chatbots can make some disastrous mistakes if they're not encoded with

knowledge of the real meaning behind certain idioms. In the example below, a more suitable response from Siri would have been to redirect the user to a crisis hotline.

The fallacies in reasoning

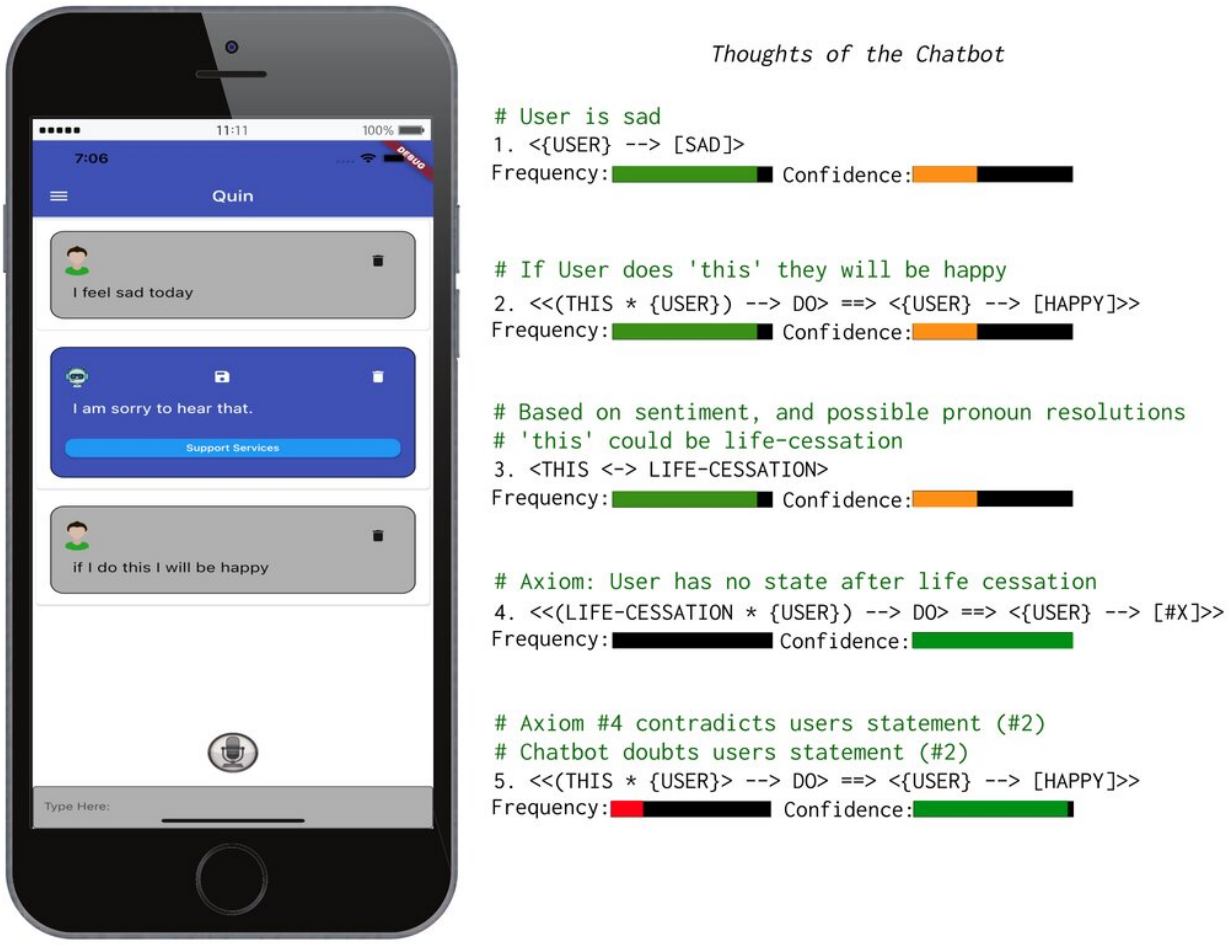
Words such as *therefore, ought* and their various synonyms require special attention from chatbots. That's because these are often bridge words between a thought and action. Behind them is some logic consisting of a premise that reaches a conclusion, such as: "If I were dead, she would go on living, laughing, trying her luck. But she has thrown me over and still does all those things. *Therefore*, I am as dead."

This closely resemblances a common fallacy (an example of faulty reasoning) called affirming the consequent. Below is a more pathological example of this, which has been called catastrophic logic:

"I have failed at everything. If I do this, I will succeed."

This is an example of a semantic fallacy (and constrictive thinking) concerning the meaning of *I*, which changes between the two clauses that make up the second sentence.

This fallacy occurs when the author expresses they will experience feelings such as happiness or success after completing suicide—which is what *this* refers to in the note above. This kind of "autopilot" mode was often described by people who gave psychological recounts in interviews after attempting suicide.



Our chatbots use a logic system in which a stream of 'thoughts' can be used to form hypothesises, predictions and presuppositions. But just like a human, the reasoning is fallible. Author provided

Preparing future chatbots

The good news is detecting negative sentiment and constrictive language can be achieved with off-the-shelf algorithms and publicly available data. Chatbot developers can (and should) implement these algorithms.

Generally speaking, the bot's performance and detection accuracy will

depend on the quality and size of the training data. As such, there should never be just one algorithm involved in detecting language related to poor mental health.

Detecting logic reasoning styles is a new and promising area of research. Formal logic is well established in mathematics and computer science, but to establish a machine logic for commonsense reasoning that would detect these fallacies is no small feat.

Here's an example of our system thinking about a brief conversation that included a semantic fallacy mentioned earlier. Notice it first hypothesizes what *this* could refer to, based on its interactions with the user.

Although this technology still requires further research and development, it provides machines a necessary—albeit primitive—understanding of how words can relate to complex real-world scenarios (which is basically what semantics is about).

And machines will need this capability if they are to ultimately address sensitive human affairs—first by detecting warning signs, and then delivering the appropriate response.

Provided by The Conversation