How to feed and raise a Wikipedia robo-editor

December 11 2015, by Mehrnoosh Sadrzadeh

'Need more input.' Credit: robo-editing, CC BY-SA

Wikipedia is to put artificial intelligence to the enormous task of keeping the free, editable online encyclopedia up-to-date, spam-free and legal. The [Objective Revision Evaluation Service](#) uses text-processing AI algorithms to scan recent edits for signs that they may be spam, trolling, revert wars (where edits are made and reversed endlessly), or otherwise dubious. But humans are excellent at making sense of the nuance of the written word – can a computer do the same?

Natural [language](#) processing is a branch of AI, focusing not on creating smart computers but on intelligent comprehension of text. Its aim is to help computers understand [human](#) language, and communicate as humans do.

Intelligent comprehension of language might mean lots of things. It might mean understanding the grammar of a language. For a computer to do this the language's internal rules must be formalised in ways a computer can understand. This is not very difficult, since grammar is a set of rules and machines are good at rule processing. Things become much more difficult with day-to-day conversations, which consist of unfinished or non-grammatical utterences such as "Well, I was going to … erm … today maybe …", or noises such as "aha", "um", "oh", "wow", which while nonsensical can nevertheless mean something to a human listener.

Understanding language might also mean being able to generate text in human ways, such as writing a novel, play, or news article. Deep neural networks have been used to train algorithms that can generate text that is similar, linguistically speaking, to the input data. An entertaining example is an algorithm that generates text [in the style of the Kings James Bible](#). Another is creating narratives based on factual data, such as a weather forecast based on temperature and winds information.

Understanding language might also mean being able to process text in

ways humans do, such as summarising, classification, paraphrasing and so on. This is what Wikipedia's robo-editors are doing, classifying edits into the real and unreal, correct and incorrect, acceptable and unacceptable.

Feeding algorithms by hand

To do any of these tasks properly, an AI must learn how to assign meaning to symbols such as words and phrases. This is a very difficult task, not least because we're not even sure how humans do it, and if we did the structure of the brain is so complex that implementing it with a computer would be even harder.

For instance research has revealed that humans are no better than chance at identifying deceptive reviews left on Trip Advisor. However, computers correctly spotted deceptive reviews 90% of the time. But this result relied on human experts to produce enough "gold standard" material – that is, truthful and fake opinions written by humans. The challenge then becomes to get hold of this training data, and the nature of the task at Wikipedia means that there isn't enough genuine, trusted data available.

In the absence of large quantities of good data, the AI needs to be trained manually, by feeding it linguistic features that can be used to distinguish the good from the bad. Psycholinguistic studies of deception have found the types of words a liar is more likely to use, for example one study found fewer causal words and negations such as "because", "effect", "no" or "never", while another study found liars avoid the use of first person pronouns (I, me, mine), but use more third person pronouns (he, she, they).

The problem is that there is a vast number of different linguistic features that could apply, and no way of knowing when one has them all – in fact

new studies are continually revealing new classes of identifying linguistic features. And some genuine texts may contain these characteristics – the robo-editor will have to work out what are the distinctive characteristics of malicious edits to Wikipedia.

However, machines are good at learning the syntax (the rules and processes) and the lexicon (the inventory of words), but do less well at modelling meaning, or "semantics". What does the robo-editor do with Wikipedia edits that are malicious, yet do not conform to the list of characteristics it has learned as representing malicious writing? How can computers understand the complexities of idioms, cynicism, metaphor and simile? It's very difficult for an algorithm to make sense of a bad edit that includes these features, or to distinguish them from valid edits that also contain them.

Despite all these challenges natural language processing is getting better and better at understanding language and performing language tasks automatically, as demonstrated by the incredible improvement in translation and intelligent search engines – those that understand what you mean, not just what you typed. Given enough data and the means to create more, AI can gradually be trained – just as human children are – to learn all aspects of human language.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation