An Algorithm Summarizes Lengthy Text Surprisingly Well

Training software to accurately sum up information in documents could have great impact in many fields, such as medicine, law, and scientific research.

May 12, 2017

Who has time to read every article they see shared on Twitter or Facebook, or every document that’s relevant to their job? As information overload grows ever worse, computers may become our only hope for handling a growing deluge of documents. And it may become routine to rely on a machine to analyze and paraphrase articles, research papers, and other text for you.

An algorithm developed by researchers at Salesforce shows how computers may eventually take on the job of summarizing documents. It uses several machine-learning tricks to produce surprisingly coherent and accurate snippets of text from longer pieces. And while it isn’t yet as good as a person, it hints at how condensing text could eventually become automated.

The algorithm produced, for instance, the following summary of a recent New York Times article about Facebook trying to combat fake news ahead of the U.K.’s upcoming election:

- Social network published a series of advertisements in newspapers in Britain on Monday.
- It has removed tens of thousands of fake accounts in Britain.
- It also said it would hire 3,000 more moderators, almost doubling the number of people worldwide who scan for inappropriate or offensive content.

The Salesforce algorithm is dramatically better than anything developed previously, according to a common software tool for measuring the accuracy of text summaries.

“I don’t think I’ve ever seen such a large improvement in any [natural-language-processing] task,” says Richard Socher, chief scientist at

https://www.technologyreview.com/s/607828/an-algorithm-s...
Salesforce. Socher is a prominent name in machine learning and natural-language processing, and his startup, MetaMind, was acquired by Salesforce in 2016.

The software is still a long way from matching a human's ability to capture the essence of document text, and other summaries it produces are sloppier and less coherent. Indeed, summarizing text perfectly would require genuine intelligence, including commonsense knowledge and a mastery of language.

Caiming Xiong, a research scientist at Salesforce who contributed to the work, says his team's algorithm, while imperfect, could summarize daily news articles, or provide a synopsis of customer e-mails. The latter could be especially useful for Salesforce's own platform.

The team's algorithm uses a combination of approaches to achieve its improvement. The system learns from examples of good summaries, an approach called supervised learning, but also employs a kind of artificial attention to the text it is ingesting and outputting. This helps ensure