**STEM Friday #13**

Prompt: Read the following article and write 2 pages about how artificial intelligence machines reading would be beneficial and why this invention matters.

Things to think about: people with disabilities, blind people, people of different languages (could we set the machine to read in different languages), auditory learners, etc.

**Google DeepMind Teaches Artificial Intelligence Machines to Read**

The best way for AI machines to learn is by feeding them huge data sets of annotated examples, and the Daily Mail has unwittingly created one.

A revolution in artificial intelligence is currently sweeping through computer science. The technique is called deep learning and it’s affecting everything from facial and voice to fashion and economics.



But one area that has not yet benefitted is natural language processing—the ability to read a document and then answer questions about it. That’s partly because deep learning machines must first learn their trade from vast databases that are carefully annotated for the purpose. However, these simply do not exist in sufficient size to be useful.

Today, that changes thanks to the work of Karl Moritz Hermann at Google DeepMind in London and a few pals. These guys say the special way that the Daily Mail and CNN write online news articles allows them to be used in this way. And the sheer volume of articles available online creates for the first time, a database that computers can use to learn and then answer related about. In other words, DeepMind is using Daily Mail and CNN articles to teach computers to read.

The deep learning revolution has come about largely because of two breakthroughs. The first is related to neural networks, where computer scientists have developed new techniques to train networks with many layers, a task that has been tricky because of the number of parameters that must be fine-tuned. The new techniques essentially produce “ready-made” nets that are ready to learn.

But a neural network is of little use without a database to learn from. Such a database has to be carefully annotated so that the machine has a gold standard to learn from. For example, for face recognition, the training database must contain pictures in which faces and their positions in the frame are clearly identified. And so that the images cover as many facial arrangements as possible, the databases have to be huge.

That’s recently become possible thanks to crowdsourcing services like Amazon’s Mechanical Turk. Various teams have created this kind of gold standard database by showing people pictures and asking them to draw bounding boxes around the faces they contain.

But creating a similarly annotated database for the written word is much harder. Sure, it’s possible to extract sentences that contain important points. But these aren’t much help because any machine algorithm quickly learns to hunt through the text for the same phrase, a trivial task for a computer.

Instead, the annotation must describe the content of the text but without appearing within it. To understand the link, a learning algorithm must then look beyond the mere occurrence of words and phrases but also at their grammatical links and causal relationships.

Creating such a database is easier said than done. Computer scientists have generated small versions by hand but these are too tiny to be of much use to a neural network. And there seems little possibility of creating larger ones by hand because humans are generally poor at annotating text accurately, unless they are specialist editors.

Enter the Daily Mail website, MailOnline, and CNN online. These sites display news stories with the main points of the story displayed as bullet points that are written independently of the text. “Of key importance is that these summary points are abstractive and do not simply copy sentences from the documents,” say Hermann and co.

That immediately suggests a way of creating an annotated database: take the news articles as the texts and the bullet point summaries as the annotation.

The DeepMind team goes further, however. They point out that it is still possible to work out the answer to many queries using simple word search approaches.

They give the following example of a type of problem known as a Cloze query, that machine learning algorithms are often used to solve. Here, the goal is to identify X in these modified headlines from the Daily Mail: a) The hi-tech bra that helps you beat breast X; b) Could Saccharin help beat X ?; c) Can fish oils help fight prostate X ?

Hermann and co point out that a simple type of data mining algorithm called an ngram search could easily find the answer by looking for words that appear most often next to all these phrases. The answer, of course, is the word “cancer.”

To foil this type of solution, Hermann and co anonymize the dataset by replacing the actors in sentences with a generic description. An example of some original text from the Daily Mail is this: *The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.”*

An anonymized version of this text would be the following:
*The****ent381****producer allegedly struck by****ent212****will not press charges against the “****ent153****” host, his lawyer said friday.****ent212****, who hosted one of the most - watched television shows in the world, was dropped by the****ent381****wednesday after an internal investigation by the****ent180****broadcaster found he had subjected producer****ent193****“ to an unprovoked physical and verbal attack .”*

In this way it is possible to convert the following Cloze-type query to identify X from “*Producer X will not press charges against Jeremy Clarkson, his lawyer says*” to “*Producer X will not press charges against****ent212****, his lawyer says*.”

And the required answer changes from “Oisin Tymon” to “***ent212***.”

In that way, the anonymized actor is only possible to identify with some kind of understanding of the grammatical links and causal relationships between the entities in the story.

The resulting database is vast, consisting of 110,000 articles from CNN and 218,000 articles from the Daily Mail website.

Having created this kind of database for the first time, Hermann and co can’t resist using it to put several machine learning techniques through their paces. They compare conventional natural language processing techniques, such as measuring the distance between combinations of words, and more modern neural network approaches.

The results clearly show how powerful neural nets have become. Hermann and co say the best neural nets can answer 60 percent of the queries put to them. They suggest that these machines can answer all queries that are structured in a simple way and struggle only with queries that have more complex grammatical structures.

There are some caveats of course. The most obvious is that articles from the Daily Mail and CNN have a very specific underlying structure that differs from other nonjournalistic forms of writing. Just how this underlying structure influences the results isn’t clear.

Neither is it clear how these machines compare to human capabilities, something that would be straightforward to find out using services like Mechanical Turk. That would put in context DeepMind’s claim, implied in the title of its paper, that these machines are learning to comprehend what they read.

Nevertheless, this is interesting work that sets the scene for some fascinating developments in the near future. Machine reading is coming; the only question is how quickly.

Ref: [arxiv.org/abs/1506.03340](http://arxiv.org/abs/1506.03340%22%20%5Ct%20%22_blank) : Teaching Machines to Read and Comprehend