Bible helps researchers perfect translation algorithms

October 23 2018

Texts from 34 versions of the English-language Bible were used to help improve computer-based style transfer systems. The result can create different versions of written passages to suit specific audiences. Credit: Bible photo: Chris Downer. Composite illustration: Keith Carlson.

In search of inspiration for improving computer-based text translators, researchers at Dartmouth College turned to the Bible for guidance. The result is an algorithm trained on various versions of the sacred texts that can convert written works into different styles for different audiences.

Internet tools to translate text between languages like English and Spanish are widely available. Creating style translators—tools that keep text in the same language but transform the style—have been much

slower to emerge. In part, efforts to develop the translators have been stymied by the difficulty of acquiring the enormous amount of data required. This is where the research team turned to the Bible.

In addition to being a source of spiritual guidance for many people around the globe, the Dartmouth-led team saw in the Bible "a large, previously untapped dataset of aligned parallel text." Beyond providing infinite inspiration, each version of the Bible contains more than 31,000 verses that the researchers used to produce over 1.5 million unique pairings of source and target verses for machine-learning training sets.

According to the research published in the journal *Royal Society Open Science*, this is not the first parallel dataset created for style translation. But it is the first that uses the Bible. Other texts that have been used in the past, ranging from Shakespeare to Wikipedia entries, provide data sets that are either much smaller or not as well suited for the task of learning style translation.

"The English-language Bible comes in many different written styles, making it the perfect source text to work with for style translation," said Keith Carlson, a Ph.D. student at Dartmouth and lead author of the research paper about the study.

As an added benefit for the research team, the Bible is already thoroughly indexed by the consistent use of book, chapter and verse numbers. The predictable organization of the text across versions eliminates the risk of alignment errors that could be caused by automatic methods of matching different versions of the same text.

"The Bible is a 'divine' data set to work with to study this task," said Daniel Rockmore, a professor of computer science at Dartmouth and contributing author on the study. "Humans have been performing the task of organizing Bible texts for centuries, so we didn't have to put our

faith into less reliable alignment algorithms."

To define "style" for the study, the researchers reference sentence length, the use of passive or active voices, and word choice that could result in texts with varying degrees of simplicity or formality. According to the study: "Different wording may convey different levels of politeness or familiarity with the reader, display different cultural information about the writer, be easier to understand for certain populations."

The team used 34 stylistically distinct Bible versions ranging in linguistic complexity from the "King James Version" to the "Bible in Basic English." The texts were fed into two algorithms—a statistical machine translation system called "Moses" and a neural network framework commonly used in machine translation, "Seq2Seq."

While different versions of the Bible were used to train the computer code, systems could ultimately be developed that translate the style of any written text for different audiences. As example, a style translator could take an English-language selection from "Moby Dick" and translate it into different versions suitable for young readers, non-native English speakers, or any one of a variety of audiences.

"Text simplification is only one specific type of style transfer. More broadly, our systems aim to produce text with the same meaning as the original, but do so with different words," said Carlson.

Dartmouth College has a long history of innovation in computer science. The term "artificial intelligence" was coined at Dartmouth during a 1956 conference that created the AI research discipline. Other advancements include the design of BASIC—the first general-purpose and accessible programing language—and the Dartmouth Time-Sharing System that contributed to the modern day operating system.

More information: Evaluating Prose Style Transfer with the Bible, *Royal Society Open Science*, rsos.royalsocietypublishing.or … /10.1098/rsos.171920

Provided by Dartmouth College