Machines that learn language more like kids do

Computer model could improve human-machine interaction, provide insight into how children learn language.

Rob Matheson | MIT News Office

Children learn language by observing their environment, listening to the people around them, and connecting the dots between what they see and hear. Among other things, this helps children establish their language's word order, such as where subjects and verbs fall in a sentence.

In computing, learning language is the task of syntactic and semantic parsers. These systems are trained on sentences annotated by humans that describe the structure and meaning behind words. Parsers are becoming increasingly important for web searches, natural-language database querying, and voice-recognition systems such as Alexa and Siri. Soon, they may also be used for home robotics.

But gathering the annotation data can be time-consuming and difficult for less common languages. Additionally, humans don't always agree on the annotations, and the annotations themselves may not accurately reflect how people naturally speak.

In a paper being presented at this week's Empirical Methods in Natural Language Processing conference, MIT researchers describe a parser that learns through observation to more closely mimic a child's language-acquisition process, which could greatly extend the parser's capabilities. To learn the structure of language, the parser observes captioned videos, with no other information, and associates the words with recorded objects and actions. Given a new sentence, the parser can then use what it's learned about the structure of the language to accurately predict a sentence's meaning, without the video.

This "weakly supervised" approach — meaning it requires limited training data — mimics how children can observe the world around them and learn language, without anyone providing direct context. The approach could expand the types of data and reduce the effort needed for training parsers, according to the researchers. A few directly annotated sentences, for instance, could be combined with many captioned videos, which are easier to come by, to improve performance.

In the future, the parser could be used to improve natural interaction between humans and personal robots. A robot equipped with the parser, for instance, could constantly observe its environment to reinforce its understanding of spoken commands, including when the spoken sentences aren't fully grammatical or clear. "People talk to each other in partial sentences, run-on thoughts, and jumbled language. You want a robot in your home that will adapt to their particular way of speaking … and still figure out what they mean," says co-author Andrei Barbu, a researcher in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Center for Brains, Minds, and Machines (CBMM) within MIT's McGovern Institute.

The parser could also help researchers better understand how young children learn language. "A child has access to redundant, complementary information from different modalities, including hearing parents and siblings talk about the world, as well as tactile information and visual information, [which help him or her] to understand the world," says co-author Boris Katz, a principal research scientist and head of the InfoLab Group at CSAIL. "It's an amazing puzzle, to process all this simultaneous sensory input. This work is part of bigger piece to understand how this kind of learning happens in the world."

Co-authors on the paper are: first author Candace Ross, a graduate student in the Department of Electrical Engineering and Computer Science and CSAIL, and a researcher in CBMM; Yevgeni Berzak PhD '17, a postdoc in the Computational Psycholinguistics Group in the Department of Brain and Cognitive Sciences; and CSAIL graduate student Battushig Myanganbayar.

Visual learner

For their work, the researchers combined a semantic parser with a computer-vision component trained in object, human, and activity recognition in video. Semantic parsers are generally trained on sentences annotated with code that ascribes meaning to each word and the relationships between the words. Some have been trained on still images or computer simulations.

The new parser is the first to be trained using video, Ross says. In part, videos are more useful in reducing ambiguity. If the parser is unsure about, say, an action or object in a sentence, it can reference the video to clear things up. "There are temporal components — objects interacting with each other and with people — and high-level properties you wouldn't see in a still image or just in language," Ross says.

The researchers compiled a dataset of about 400 videos depicting people carrying out a number of actions, including picking up an object or putting it down, and walking toward an object. Participants on the crowdsourcing platform Mechanical Turk then provided 1,200 captions for those videos. They set aside 840 video-caption examples for training and tuning, and used 360 for testing. One advantage of using vision-based parsing is "you don't need nearly as much data — although if you had [the data], you could scale up to huge datasets," Barbu says.

In training, the researchers gave the parser the objective of determining whether a sentence accurately describes a given video. They fed the parser a video and matching caption. The parser extracts possible meanings of the caption as logical mathematical expressions. The sentence, "The woman is picking up an apple," for instance, may be expressed as: *λxy.* woman *x,* pick_up *x y,* apple *y.*

Those expressions and the video are inputted to the computer-vision algorithm, called "Sentence Tracker," developed by Barbu and other researchers. The algorithm looks at each video frame to track how objects and people transform over time, to determine if actions are playing out as described. In this way, it determines if the meaning is possibly true of the video.

Connecting the dots

The expression with the most closely matching representations for objects, humans, and actions becomes the most likely meaning of the caption. The expression, initially, may refer to many different objects and actions in the video, but the set of possible meanings serves as a training signal that helps the parser continuously winnow down possibilities. "By assuming that all of the sentences must follow the same rules, that they all come from the same language, and seeing many captioned videos, you can narrow down the meanings further," Barbu says.

In short, the parser learns through passive observation: To determine if a caption is true of a video, the parser by necessity must identify the highest probability meaning of the caption. "The only way to figure out if the sentence is true of a video [is] to go through this intermediate step of, 'What does the sentence mean?' Otherwise, you have no idea how to connect the two," Barbu explains. "We don't give the system the meaning for the sentence. We say, 'There's a sentence and a video. The sentence has to be true of the video. Figure out some intermediate representation that makes it true of the video.'"

The training produces a syntactic and semantic grammar for the words it’s learned. Given a new sentence, the parser no longer requires videos, but leverages its grammar and lexicon to determine sentence structure and meaning.

Ultimately, this process is learning “as if you’re a kid,” Barbu says. “You see world around you and hear people speaking to learn meaning. One day, I can give you a sentence and ask what it means and, even without a visual, you know the meaning.”

“This research is exactly the right direction for natural language processing,” says Stefanie Tellex, a professor of computer science at Brown University who focuses on helping robots use natural language to communicate with humans. “To interpret grounded language, we need semantic representations, but it is not practicable to make it available at training time. Instead, this work captures representations of compositional structure using context from captioned videos. This is the paper I have been waiting for!”

In future work, the researchers are interested in modeling interactions, not just passive observations. “Children interact with the environment as they’re learning. Our idea is to have a model that would also use perception to learn,” Ross says.

This work was supported, in part, by the CBMM, the National Science Foundation, a Ford Foundation Graduate Research Fellowship, the Toyota Research Institute, and the MIT-IBM Brain-Inspired Multimedia Comprehension project.