An evaluation metric for determining if a chatbot is just chatty, or engaging

July 10 2020

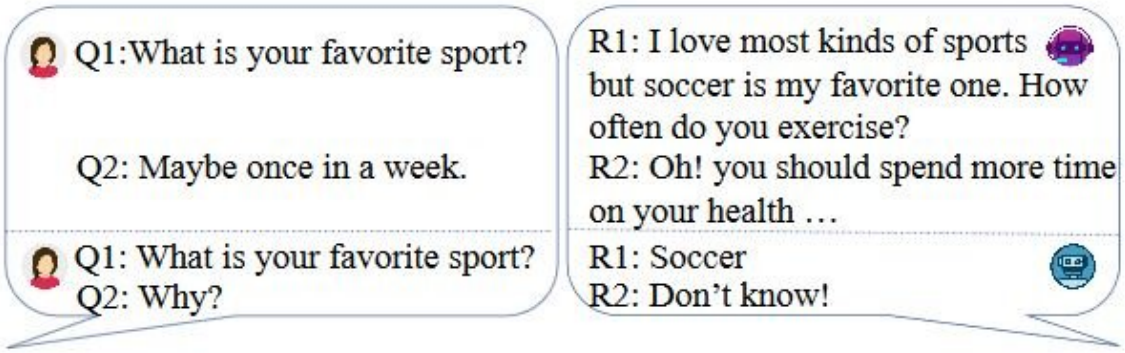


Figure 1: An illustrative example of a user's conversation with two chatbots. We anticipate that the user will prefer to converse with the top chatbot because the responses are both relevant and engaging, while the bottom chatbot generates relevant but not engaging responses.

The team's research emphasizes that more than just giving relevant responses, a chatbot must be engagin, as well. Credit: University of Southern California

From purchases to therapy to friendship, it seems as though theres a chatbot for just about everything.

And now, researchers at the USC Viterbi School of Engineering's

Information Sciences Institute (ISI) have a new way to grade the conversational skills of Cleverbot or Google's Meena. It's one thing to provide short, helpful answers to questions like, "What time does my flight leave tomorrow?" But in a world where chatbots are increasingly called upon to respond to therapeutic questions or even engage as friends, can they hold our attention for a longer conversation?

In a paper presented at the 2020 AAAI Conference on Artificial Intelligence, USC researchers announced a new evaluation metric, "predictive engagement," which rates chatbot responses on an engagement scale of "0-1," with a "1" being the most engaging.

The metric was developed by lead author and Ph.D. student Sarik Ghazarian and advised by Aram Galstyan, USC ISI director of AI and principal scientist, and Nanyun Peng, USC Viterbi research assistant professor. Ralph Weischedel, senior supervising computer scientist and research team leader, served as PI on the project.

Improved Evaluation = Improved Engagement

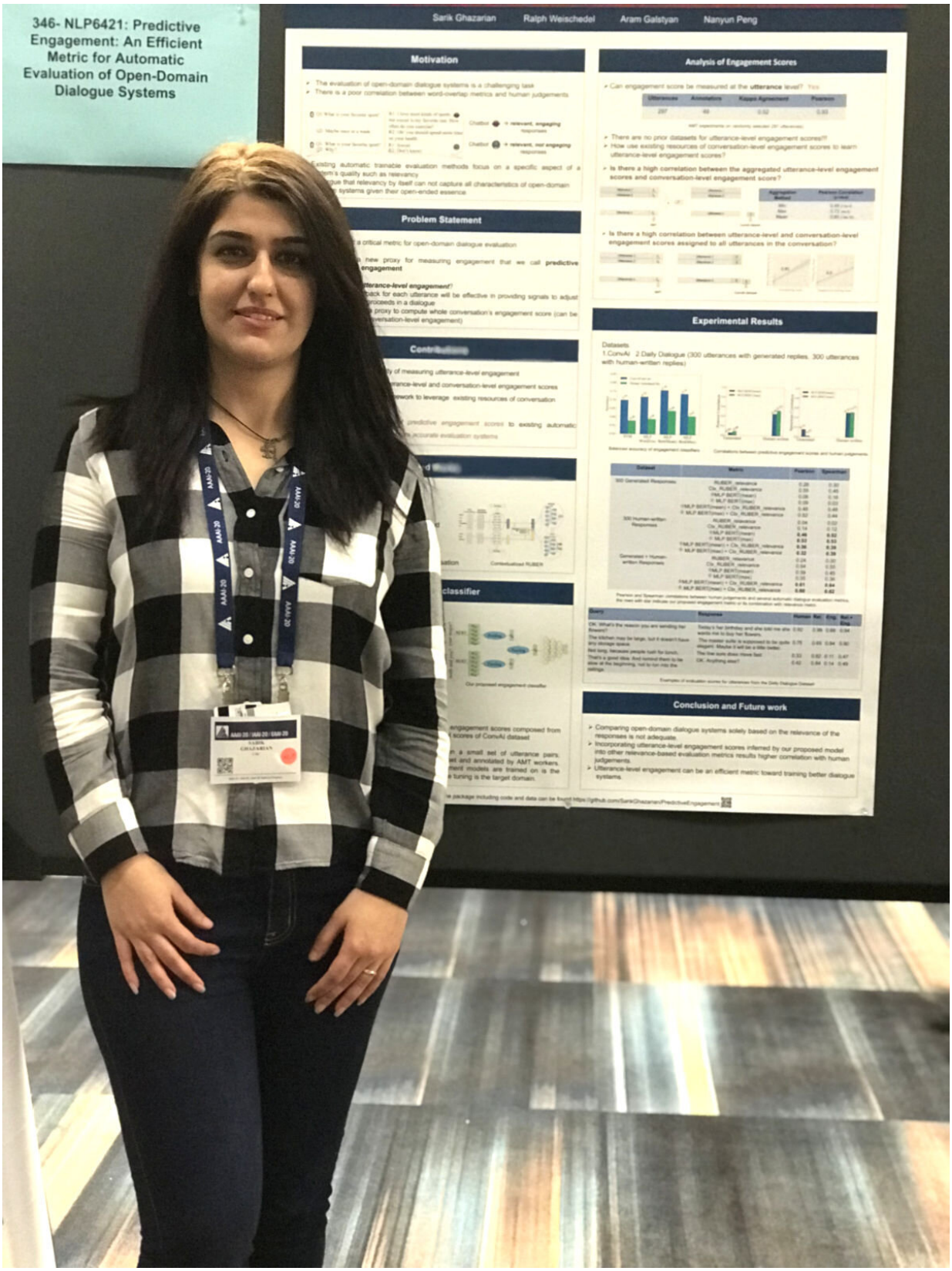
So why the need for a new metric in the first place? As Ghazarian explained, it's much more difficult to evaluate how well something like a chatbot is conversing with a user, since a chatbot can be an open-domain dialog system through which the interaction mostly contains open-ended information.

A dialog system is essentially a computer system that incorporates text, speech, and other gestures in order to converse with humans. There are two general types: task-oriented dialog systems are useful when we want to achieve a specific goal, such as reserving a room in a hotel, purchasing a ticket, or booking a flight. Open-domain dialog systems, on the other hand, such as chatbots, focus more on interacting with people on a deeper level, and they do so by imitating human-to-human conversations.

"The evaluation of open-domain dialog systems is one of the most substantial phases in the development of high-quality systems," she said. "In comparison to task-oriented dialogs, [where] the user converses to achieve a predetermined goal, the evaluation of open-domain dialog systems is more challenging. The user who converses with the open-domain dialog systems doesn't follow any specific goal, [so] the evaluation can't be measured on whether or not the user has achieved the purpose."

In their paper, the ISI researchers underlined that evaluating open-domain dialog systems shouldn't be restricted to only specific aspects, such as relevancy-the responses also need to be genuinely interesting to the user.

"The responses generated by an open-domain dialog system are admissible when they're relevant to users and also interesting," Ghazarian continued. "We showed that incorporating the interestingness aspect of responses, which we call as engagement score, can be very beneficial toward having a more accurate evaluation of open-domain dialog systems." And, as Ghazarian noted, understanding the evaluation will help improve chatbots and other open-domain dialog systems. "We plan to explore the effects of our proposed automated engagement metric on the training of better open-domain dialog systems," she added.


346- NLP6421: Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems
Sarik Ghazarian
Ralph Weischedel
Aram Galstyan
Nanyun Peng
Motivation
Problem Statement
Analysis of Engagement Scores
Experimental Results
Conclusion and Future work

Sarik Ghazarian presented the predictive engagement metric research at AAAI 2020. Credit: University of Southern California

Alexa, Play "Computer Love"

Chatbots such as Cleverbot, Meena, and XiaoIce are able to engage people in exchanges that are more akin to real life discussions than task-oriented dialog systems.

XiaoIce, Microsoft's chatbot for 660 million Chinese users, for example, has a personality that simulates an intelligent teenage girl, and along with providing basic AI assistant functions, she can also compose original songs and poetry, play games, read stories, and even crack jokes. XiaoIce is described as an "empathetic chatbot," as it attempts to develop a connection and create a friendship with the human it's interacting with.

"[These types of chatbots] can be beneficial for people who are not socialized so that they can learn how to communicate in order to make new friends," said Ghazarian.

Open-domain chatbots-ones that engage humans on a much deeper level-are not only gaining prevalence, but are also getting more advanced. "Even though task-oriented chatbots are most well-known for their vast usages in daily life, such as booking flights and supporting customers, their open-domain counterparts also have very extensive and critical applications that shouldn't be ignored," Ghazarian remarked, underlining that the main intention of a user's interaction with these types of chatbots isn't only for entertainment but also for general knowledge.

For example, open-domain chatbots can be utilized for more serious

issues. "Some of these chatbots are designed to [provide] mental health support for people who face depression or anxiety," Ghazarian explained. "Patients can leverage these systems to have free consultations whenever they need them." She pointed to a study funded by the U.S. Defense Advanced Research Projects Agency (DARPA), which found that people find it easier to talk about their feelings and personal problems when they know they're conversing with a chatbot, as they feel it won't judge them.

Therapy chatbots such as Woebot have been found to be effective at implementing real-life therapeutic methods, such as cognitive behavioral therapy, and certain studies have indicated that users of Woebot may have an easier time even outside of therapy sessions, since they have 24/7 access to the chatbot. Additionally, chatbots may also be helpful in encouraging people to seek treatment earlier. For instance, Wysa, a chatbot marketed as an "AI life coach," has a mood tracker than can detect when the user's mood is low and offers a test to analyze how depressed the user is, recommending professional assistance as appropriate based on the results.

Open-domain chatbots are also extremely useful for people that are learning a foreign language. "In this scenario, the chatbot plays the role of a person who can talk the foreign language and it tries to simulate a real conversation with the user anytime and anywhere," said Ghazarian. "This is specifically useful for people who don't have confidence in their language skills or even are very shy to contact real people."

The predictive engagement metric will help researchers better evaluate these types of chatbots, as well open-domain dialog systems overall. "There are several applications of this work: first, we can use it as a development tool to easily automatically evaluate our systems with low cost," Peng said. "Also, we can explore integrating this evaluation score as feedback into the generation process via reinforcement learning to

improve the dialog system."

By better understanding how the evaluation works, researchers will be able to improve the system itself. "Evaluating open-domain natural language generation (including dialog systems) is an open challenge," Peng continued. "There are some efforts on developing automatic evaluation metrics for open-domain dialog systems, but to make them really useful, we still need to push the correlation between the automatic evaluation and human judgment higher-that's what we've been doing."

More information: Ghazarian et al., Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems. arXiv:1911.01456 [cs.CL]. arxiv.org/pdf/1911.01456.pdf

Provided by University of Southern California