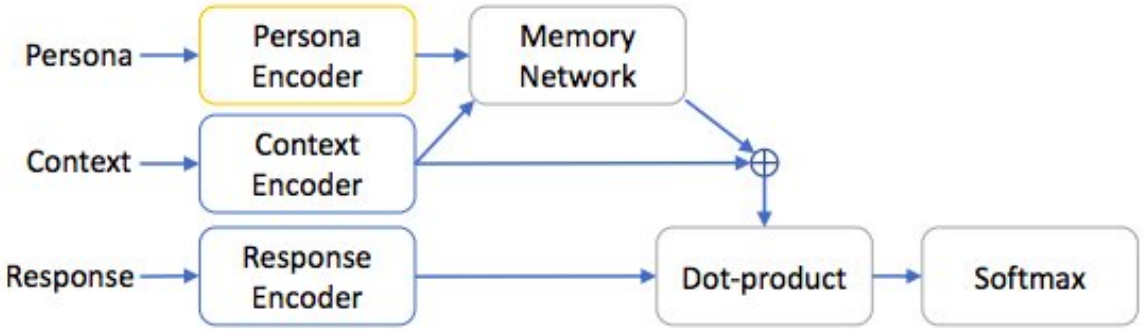
Facebook researchers build a dataset to train personalized dialogue agents

September 18 2018, by Ingrid Fadelli



Persona-based network architecture. Credit: Mazaré et al.

Researchers at Facebook have recently compiled a dataset of 5 million personas and 700 million persona-based dialogues. This database could be used to train end-to-end dialogue systems, resulting in more engaging and rich dialogues between computer agents and humans.

Dialogue systems, or conversational agents (CA), are computer systems designed to communicate with human beings via text, speech, graphics, or other methods, in a coherent way. So far, dialogue systems based on neural architectures, such as LSTMs or memory networks, have been found to be particularly promising in achieving fluent communication,

particularly when trained directly on dialogue logs.

"One of their main advantages is that they can rely on large data sources of existing dialogues to learn to cover various domains without requiring any expert knowledge," the researchers wrote in their paper, which was pre-published on arXiv. "However, the flip side is that they also exhibit limited engagement, especially in chit-chat settings: They lack consistency and do not leverage proactive engagement strategies as (even partially) scripted chatbots do."

In a recent study, a different team of researchers at Montreal Institute for Learning Algorithms (MILA) and Facebook AI created a dataset called PERSONA-CHAT, which includes dialogues between agents with text profiles, or personas, attached to them. They found that training a dialogue system on a particular persona improved their engagement in interactions.

"However, the PERSONA-CHAT dataset was created using an artificial data collection mechanism based on Mechanical Turk," the researchers explained in their paper. "As a result, neither dialogs nor personas can be fully representative of real user-bot interactions and the dataset coverage remains limited, containing a bit more than 1k different personas."

To address the limitations of the previously compiled dataset, the Facebook researchers created a new, large-scale persona-based dialogue dataset, composed of conversations extracted from online platform Reddit. Their study takes the work of their predecessors one step further, by using more representative interactions.

"In this paper, we build a very large-scale persona-based dialogue dataset using conversations previously extracted from Reddit," the researchers wrote. "With simple heuristics, we create a corpus of over 5 million personas spanning more than 700 million conversations."

To evaluate its effectiveness, the researchers trained persona-based end-to-end dialogue systems on their newly developed dataset. Systems trained on their dataset were able to conduct more engaging conversations, outperforming other conversational agents that did not have access to personas during their training.

Interestingly, their dataset led to state-of-the-art results even when dialogue systems were merely pre-trained on it. In future, these findings could lead to the development of more engaging chatbots, which can also be personalized and trained to acquire a particular persona.

"We show that training models to align answers both with the persona of their author and the context improves the predicting performance," the researchers wrote. "As pre-training leads to considerable improvement in performance, future work could fine-tune this model for various dialogue systems."

More information: Training Millions of Personalized Dialogue Agents. arXiv:1809.01984 [cs.CL]. arxiv.org/abs/1809.01984

Personalizing Dialogue Agents: I have a dog, do you have pets too? arxiv.org/abs/1801.07243