Self-Explaining for Intuitive Interaction with AI

Since it is increasingly harder to opt out from interacting with AI technology, people demand that AI is capable of maintaining contracts such that it supports agency and oversight of people who are required to use it or who are affected by it. To help those people create a mental model about how to interact with AI systems, I extend the underlying models to self-explain—predict the label/answer and explain this prediction. In this talk, I will present how to generate (1) free-text explanations given in plain English that immediately tell users the gist of the reasoning, and (2) contrastive explanations that help users understand how they could change the text to get another label.

Bio: Ana Marasović is a postdoctoral researcher at the Allen Institute for AI (AI2) and the Paul G. Allen School of Computer Science & Engineering at University of Washington. Her research interests broadly lie in the fields of natural language processing, explainable AI, and multimodal learning. Her projects are motivated by a unified goal: improve interaction and control of the NLP systems to help people make these systems do what they want with the confidence that they’re getting exactly what they need. Prior to joining AI2, Ana obtained her PhD from Heidelberg University.