“I Don’t Think So”: Summarizing Policy Disagreements for Agent Comparison

Abstract: With Artificial Intelligence on the rise, human interaction with autonomous agents becomes more frequent. Effective human-agent collaboration requires users to understand the agent’s behavior, as failing to do so may cause reduced productivity, misuse or frustration. Agent strategy summarization methods are used to describe the strategy of an agent to its destined user through demonstration. A summary’s objective is to maximize the user’s understanding of the agent’s aptitude by showcasing its behaviour in a selected set of world states. While shown to be useful, we show that current methods are limited when tasked with comparing between agents, as each summary is independently generated for a specific agent. In this paper, we propose a novel method for generating dependent and contrastive summaries that emphasize the differences between agent policies by identifying states in which the agents disagree on the best course of action. We conduct user studies to assess the usefulness of disagreement-based summaries for identifying superior agents and conveying agent differences. Results show disagreement-based summaries lead to improved user performance compared to summaries generated using HIGHLIGHTS, a strategy summarization algorithm which generates summaries for each agent independently.