Retrieval Augmented Generation for Immersive Formal Dialogue

Mark Snaith and Simon Wells

The relatively recent emergence of Large Language Models (LLMs) has dramatically changed the Artificial Intelligence landscape. In particular, they have demonstrated powerful capabilities to underpin conversational systems that ostensibly exhibit understanding. For example, platforms such as ChatGPT and Copilot provide chat-style interfaces that allow users to make requests (prompts) which are then serviced in a natural, conversational manner.

LLMs are driven by prompts: given a prompt as input, they will predict the best response to make. By default, however, LLMs have no consideration of conversational flow, and in general interactions therewith are very open and unstructured. For example, one prompt might ask an LLM for ideas about what to do on a rainy day, while the next prompt can request instructions for how to solve a puzzle. LLMs can also hallucinate, meaning they present as fact statement that are either untrue, or at least unverifiable.

In contrast to the open and unstructured nature of LLMs, formal dialogue games provide a structured account of how a multi-party, goal-oriented conversation should proceed with little or no consideration of the precise content of each move. These models have subsequently been used to underpin human-agent interactions in computational applications across a variety of domains. However, these applications have historically encountered a human-level vocabulary gap, resulting in agent responses that are not as natural as those found in human-human conversations.

There is therefore an opportunity to address these dual issues: using formal dialogue games to regulate LLM utterances in goal-oriented conversations, and using LLMs to bridge the human-level vocabulary gap found in dialogue games. One aspect of addressing these issues to ensure that LLM responses make use only of domain-specific knowledge relevant to the conversation at hand.

Retrieval Augmented Generation (RAG) is a technique that allows LLMs to call upon external sources for all or part of their response generation. This allows for the development of systems that harness language generation capabilities of LLMs, while ensuring those generated responses are grounded in the provided knowledge. By ensuring that the LLM uses only the knowledge provided, RAG-based approaches have proven highly effective in overcoming some of the drawbacks of LLMs, including hallucinations and biases.

We present in this paper a pipeline for Retrieval Augmented Generation (RAG) to support computational accounts of formal dialogue. Given a possible

dialogue move type, the pipeline harnesses the generative capabilities of LLMs to create a natural-sounding response that move type, grounded in a variety of provided knowledge sources. The pipeline has been designed to support a variety of knowledge sources, both structured (e.g. argument analyses) and unstructured (e.g. natural language documents), while also allowing developers to incorporate their own, possibly domain-specific, knowledge representations.