SIMON WELLS & MARK SNAITH

ON THE ROLE OF DIALOGUE MODELS IN THE AGE OF LLMS (SOME PRELIMINARY INVESTIGATIONS)

Presented to the 23rd International Workshop on Computational Models of Natural Argument held online on 1st December 2023

WHY FOCUS ON THIS?

- Both authors have worked with/on (formal) dialogue, i.e. dialogue/dialectical game for many years
  - Simon: Since UG studies (implemented PPD games from Walton & Krabbe (1995). Then Ph.D studies lead to the Dialogue Game Description Language [Wells (2007), Wells & Reed (2012)]
  - Mark: Since PhD studies - Applying dialogue games, through DGDL & DGEP, to real world problems, e.g. coaching dialogues [Snaith *et al* (2018, 2019)]
  - Vested (conflict of) interest in ongoing research in this area
- ML-based approaches to NLP have made great gains in recent years. Particularly LLMs are increasingly capable of generating *plausible* responses to prompts. Chatbots, utilising LLMs to respond to user input are increasingly adopted.
- Led to ask: what's the role of dialogue models in this new "AI" world?

DIALOGUE MODELS

- Formal Dialectical Games (although there are also informal descriptions of dialogue that usefully describe human behaviour, i.e. Pragma Dialectics)
  - Multi player, turn-taking, games in which players take turns to make moves where the moves are locutions (utterances containing some combination of explicit speech act and content)
  - Rules specify who can say what and when it can be said in light of what has already been said
  - Rules specify how the dialogue begins, progresses, and (ideally) terminates
- The Dialogue Game Description Language (DGDL) [Wells & Reed (2012)] is an attempt to produce a comprehensive description language for specifying the rules of dialogue games. The aim is to produce descriptions of the rules of dialogues games in a precise, comprehensive, consistent, and executable format
- The Dialogue Game Execution Platform (DGEP) is a runtime/dialogue manager for enforcing the rules of a DGDL game played between multiple human/software agents

LLMS /ML

- ML networks trained on vast amounts of real-world data
- Noted for their ability to generate plausible text and seemingly “human” responses to prompts
- Examples include:
  - OpenAI's GPT models (e.g., GPT-3.5 and GPT-4, used in ChatGPT)
  - Google's PaLM (used in Bard)
  - Meta's LLaMa
  - Anthropic's Claude 2
  - BLOOM, Ernie 3.0 Titan, etc.
- Has given rise to a focus on conversational interfaces as well as the notion of “prompt engineering” - given the right query we can engineer the desired output

INTERACTIONS

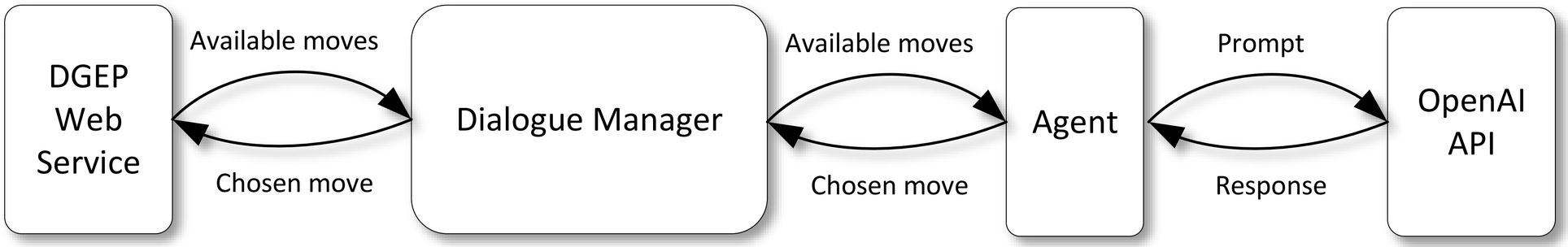
- LLM based dialogue
  - Becoming prevalent in HCI contexts, i.e. human machine “chatbot” interfaces for customer support, information search, etc.
- Formal & Informal Dialogue Models
  - Focus not just on models of actual interaction (overlapping with the LLMs) but also normative and idealised behaviours
  - Recognise that systems based purely on actual human behaviour might lead to what appear to be plausible interactions but don’t necessarily lead to systems that also matches human ideals
- LLM + DM
  - LLMs *might* be trained to interact, in extended dialogues, with people, and each other but this well take time to achieve
  - Just as we have ideals for human argument, that exceed everyday interaction, so the same ideals should exist for LLMs
  - In the meantime, dialogue models can fill in the functional gaps in achievement in LLMs, e.g. goal oriented, multi-interaction, planned, strategic dialogues. Also an additional related role in terms of providing explanation and building/maintaining trust.

EVALUATING & BENCHMARKING

- Previous approaches to evaluating dialogue, e.g McBurney & Parsons (2002) and Wells (2005) focussed on attributes of (1) the dialogue protocol and (2) the resulting dialogue
  - e.g. Stated Dialogue Purpose, Diversity of Individual Purposes, Inclusiveness, Transparency, Fairness, Clarity of Argumentation Theory, Separation of Syntax & Semantics, Rule-Consistency, Encouragement of Resolution, Discouragement of Disruption, Enablement of Self-Transformation, System Simplicity, Computational Simplicity
  - Simplicity of Representation, Efficiency of Process, Flexibility, Expressiveness, Representiveness, Stability,
- Much ML/LLM benchmarking focussed upon measures of behviour against a normative benchmark drawn from real-world human behaviour
  - e.g. Given a training dataset, does the LLM produce output that matches a specific test dataset
- In order to gain a sense of capability of, between, and across systems, we considered additional approaches to attempt to summarise what a given system (and the components making up the system) were practically capable of
- Based on addressing wh-questions posed in terms of the capability of the system (with a human providing the gold standard

DM+LLM EXPERIMENTS

- Preliminary experiments utilising simple dialogue games (expressed in DGDL & executed using DGEP)
- LLM (GPT-3.5) used to provide responses to prompts created by a simple agent. Responses used as content for the moves within a simple DGDL game managed by DGEP



SUMMARY

- We began extending evaluation approaches, stemming from dialogues games & MAS communication to account for system capabilities
  - To get a better idea of which components fulfil which responsibilities and which components are currently limited in their abilities
- We then started to investigate how simple agent dialogue systems can be built that integrate dialogue models (for strategy, planning, structure) and LLMs (initially for surface language generation but we aim to explore a creative use of prompt engineering to yield different types of dialogue move content)
- Initial results lead us to conclude that there are roles for both approaches both separately, and in combination:
  - LLMs have (increasing) capabilities in their own right
  - Dialogue games frequently focus on aspects of dialogue that are not as important, or haven't been achieved, in LLM research
  - The two approaches are complimentary and could potentially yield systems that achive more jointly than individually

DISCUSSION

- Many opportunities at the intersection of DM & ML/LLM research
  - Particularly when building on models that focus variously on normativity, ideal/perfect, or everyday behaviours
  - These help us to understand our own behaviours (humans will remain in the loop)
  - These can set benchmarks and bounds for expected, ideal, planned behaviour
  - Can inform strategic considerations - gap between the collective/statistical and the individual
  - Dialogue as a goal-oriented enterprise
- Convincing Funders that DM research should still be funded (in the age of LLMs) might remain a problem in the short/medium term
- This is preliminary and ongoing work. Things will change as both areas of research continue. It’s worth planning ahead…