Reasoning, Learning, & eXplainability Workshop (ReaLX 2018)

Towards Argumentative Dialogue as a Humane Interface between People and Intelligent Machines

Simon Wells
School of Computing
Edinburgh Napier University

simonwells.org

Edinburgh Napier UNIVERSITY

Introduction

- Trends
    - Increases in capabilities of intelligent machines
    - Increasing delegation of decision making to machines
- Issues
    - Societal need for explanation (parity of treatment)
    - Legal mandates for scrutinisation & interpretation (regulated sectors)
    - Explanation means different things to different groups
    - Acceptability of an explanation can be dependent upon target (can easily shift into justification)
    - Tendency for humans to mistrust anything different
    - Need for trust if machines are to act effectively within society

Goals

1. A system that supports explanations

2. A system that can justify decisions

3. A system that is independent of the underlying intelligent system

   ❖ Many approaches to generation/emission of explanatory artefacts [Gregor & Benbasat | Gunning | Ribeiro | Oren *et al*]

   ❖ Different forms of explanation & explanatory system/power [Doran (2018)]

   ❖ Different contexts of use of explanation (e.g. explaining to end user and engineer are different)

4. Composable

5. Work from human oriented interface towards intelligent systems (reuse what I have)

The Road to Dialogical Interaction

❖ We can model dialogues as protocols and manage interactions between speakers.

❖ Previous work:

MAgtALO - **M**ulti**A**gent **A**rgument **L**ogic & **O**pinion

[Reed & Wells (2006)]

DGDL - **D**ialogue **G**ame **D**escription **L**anguage

[Wells & Reed (2012)]

ADAMANT - **A** **D**i**A**logue **MAN**agement **T**ool

[Wells (*forthcoming*)]

Overview

1. Recognise patterns of reasoning (schemes)

2. Use schemes & NLG to instantiate arguments

3. Interact with intelligent systems via structured dialogue (explanatory & justificatory dialogues)

Argumentation Schemes

* A structure for formalising stereotypical patterns of presumptive (deductive, inductive, plausible) reasoning

* Derived from empirical studies of human reasoning & argument

* Used to catalogue, group, criticise and explore instances of reasoning

* Many 100s of identified schemes, for example:
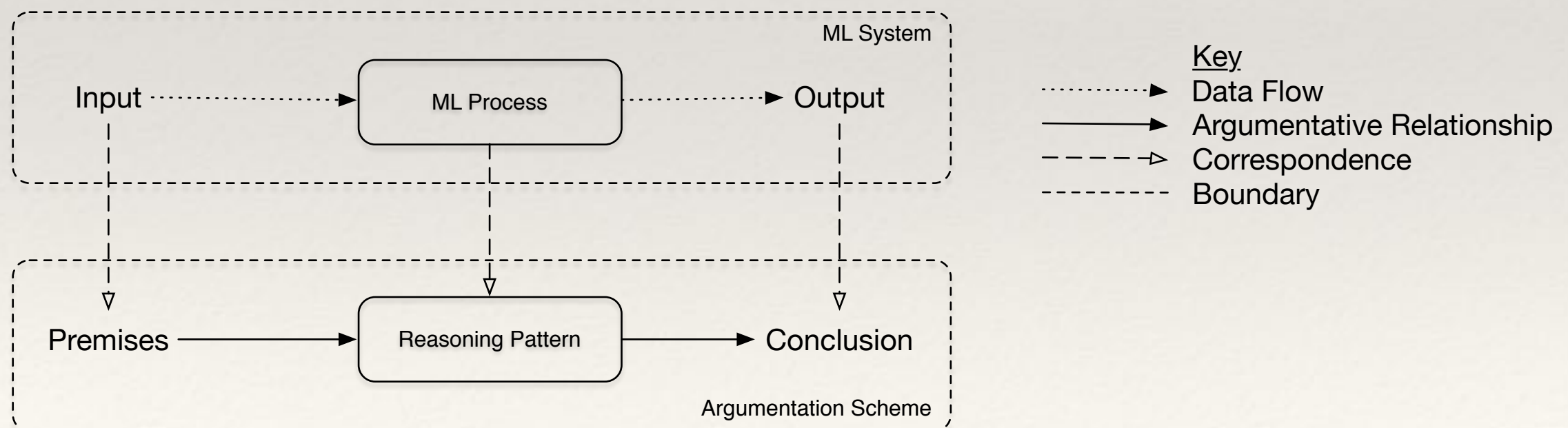
Argumentation Scheme for the Argument from Sign

Specific Premise: A (a finding) is true in this situation.
General Premise: B is generally indicated as true when its sign, A, is true.
Conclusion: B is true in this situation

Patterns of Reasoning

- Individual ML processes perform reasoning

- Nothing special (yet?) about ML reasoning

- *ergo* ML reasoning processes should map to pre-identified instances of human reasoning

- Manual mapping (OpenL, UCI ML Repo, Publications)

  - Initially: Part of the intelligent system design process

  - Future: Automated reasoning pattern (scheme) recognition?
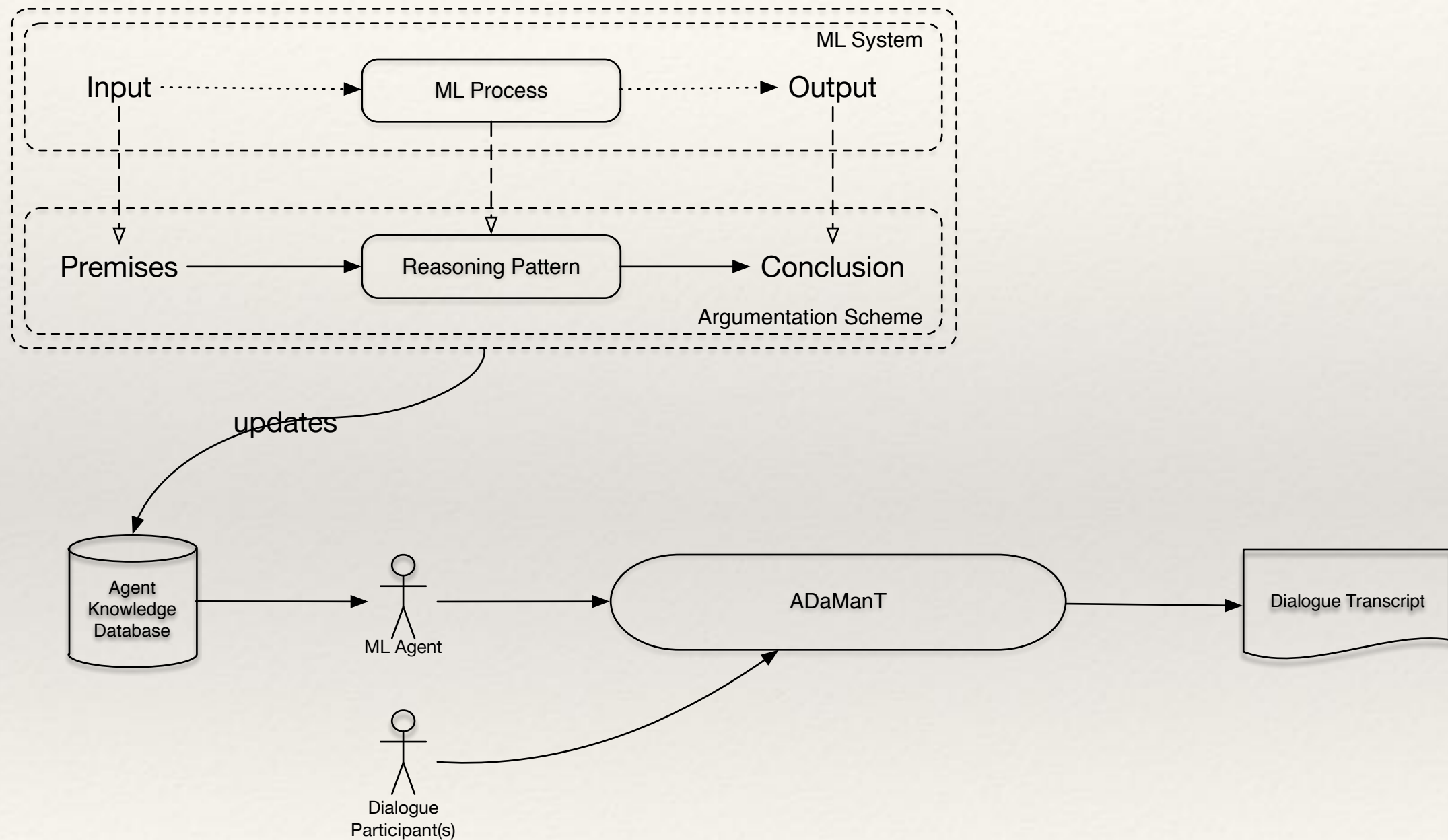
```
                                          ML System
  ┌ - - - - - - - - - - - - - - - - - - - - - - ┐
  │                ┌────────────┐                │
  │  Input ·····>  │ ML Process │  ·····> Output │
  │                └────────────┘                │
  └ - - │ - - - - - - - - │ - - - - - - │ - - - -┘
        │                 │             │
        ▽                 ▽             ▽
  ┌ - - - - - - - - - - - - - - - - - - - - - - ┐
  │              ┌──────────────────┐           │
  │  Premises ─> │ Reasoning Pattern │ ─> Conclusion
  │              └──────────────────┘           │
  │                          Argumentation Scheme│
  └ - - - - - - - - - - - - - - - - - - - - - - ┘
```

Key
········> Data Flow
─────────> Argumentative Relationship
------ -▷ Correspondence
--------- Boundary

NLG from Schemes

- Iterative process:

    - Initially using whole sentence mapping

        - (*a la* MAgALO)

    - Moving towards NLG completion of templates

    - Eventually full NLG

System Overview

Challenges

- ❖ Huge research challenges:

  - ❖ Data to knowledge/Neural-symbolic computing (Besold & Doran @ city)

  - ❖ Natural language generation (& NLU)

  - ❖ Strategic & Contextual Personalisation

Benefits

❖ A system that supports explanations

❖ A system that can justify decisions

❖ A system that is independent of the underlying intelligent system

❖ Can be used to build trust:

    ❖ I ask for a decision, then interrogate that decision and come to understand it. I get rid of the unknowns

❖ Other contexts: Legal & regulatory interaction

Conclusions

❖ Many route to get to explainable intelligent systems

❖ The human-machine interface (dialogue) is important to the relationship between people & intelligent systems

❖ There is lots of work to build on

❖ Huge challenges

References

- Doran *et al* (2018) "What Does Explainable AI Really Mean?"

- Reed & Wells (2007) "Dialogical Argument as an Interface to Complex Debates"

- Wells & Reed (2012) "A domain specific language for describing diverse systems of dialogue "

- Wells (forthcoming) "The Dialogue Game Description Language: Syntax, Semantics & Tooling"