DIALOG: Natural Language-based Interaction with a Mathemtics Assistance System

Christoph Benzmüller

Department of Computer Science, Saarland University

ΩMEGA Group Meeting

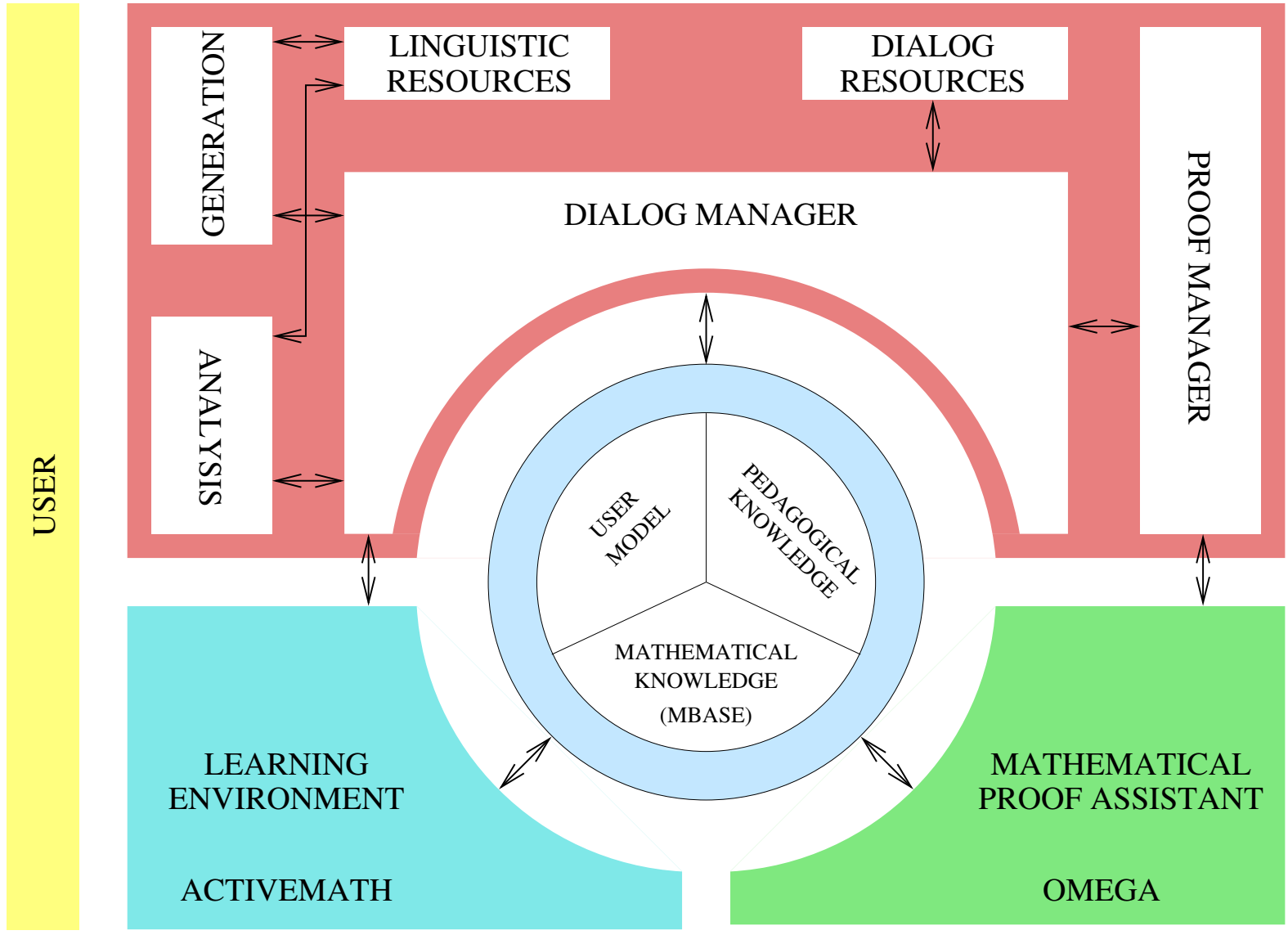
Saarbrücken, Germany

UNIVERSITÄT DES SAARLANDES

Current Project Phase

- We made a first start on opening up a new field of natural-language based mathematical tutoring dialogs.
- The foremost aim was to obtain a general view of the interplay between advanced natural language processing in a flexible tutoring dialog, and dynamic, abstract level mathematical domain reasoning.
- We moved from collecting empirical data through modeling of the different components and their interfaces to a demonstrator implementation.

Architecture



Achievements

- Experiment design, empirical test environment, and Wizard of Oz tool.
- First experiment in naive set theory domain, with written dialog input and output.
- Preliminary corpus investigation and subsequent formal annotation at several levels of interpretation: deep semantic structure, dialog moves, and tutorial task aspects.
- Coarse grained architecture and specification of refined modules for input analysis, proof management, and tutorial dialog moves (especially hinting).

Achievements (contd.)

- Realization of input analysis using a deep dependency-based grammar, focusing on uniform interpretation of informal interleaved natural language and mathematical formulae.
- Realization of the proof manager: proof representation languages for the proof manager, interfacing to the underlying domain reasoner (the Omega theorem prover); agent-based assertion reasoning that can enumerate proof step suggestions.
- Development of a demonstrator (in progress at the time of writing; to be completed before the review meeting).

Goals of Next Phase

Three important research questions that have grown out of our research:

I Processing *informal* input that consists of *interleaved* natural language text and mathematical expressions.

- We start with our framework for deep semantically-oriented analysis developed in the first project phase
- We shall *combine* the *deep analysis* approach from the first project phase *with methods for flat and partial interpretation*.
- We will extend the analysis methodology to simultaneous written and spoken input, combined also with mouse pointing.

Goals of Next Phase (contd.)

II Criteria and methods for *proof step evaluation*.

- Linguistic analysis yields a formal representation of a proof step as proposed by the user; this needs to be evaluated in tutorial context.
- Evaluation of *soundness* is straightforward ('yes' or 'no' results independent of the tutorial context)
- Evaluating the appropriateness of a proposed proof step, viz., the right *granularity* and *relevance*, is much more demanding.
- The system must be able to discriminate between target-oriented proof-step proposals and all the rest of syntactically correct and logically sound, but irrelevant, tautological, or misleading steps.

Goals of Next Phase (contd.)

III *Ambiguity* pervades all levels of processing:

- Continue work on the representation of ambiguous input.
- Tackle the new task of how to resolve these ambiguities in the given context.
- There are various linguistic sources of ambiguity, as well as utterances which are linguistically unambiguous but nevertheless do not provide enough information for an unambiguous mapping to a domain-specific interpretation.
- Challenge is *ambiguity resolution* based on domain knowledge.
- We want to explore how the CHORUS techniques to reduce or resolve ambiguity without complete enumeration of readings can be applied.

Work Packages

WP1: Interpretation of Informal Mathematical Input

WP2: Proof Management and Proof-Step Evaluation

WP3: Domain Reasoning for Ambiguity Resolution

WP4: Integration and Evaluation

Proof Step Evaluation

- Soundness:
  Can the proof step be verified by a formal inference system?
- Granularity:
  Is the granularity (i.e., 'logical size' or 'argumentative complexity') of the proof step acceptable?
- Relevance:
  Is the proof step needed or useful in achieving the goal?

Example Scenario

Assertions already introduced

(A1) $A \wedge B$.

(A2) $A \Rightarrow C$.

(A3) $C \Rightarrow D$.

(A4) $F \Rightarrow B$.

(G) $D \vee E$.

Alternative proof step directives.

(a) Aus den Annahmen folgt D.

(b) B gilt.

(c) Es genügt D zu zeigen.

(d) Wir zeigen E.

Soundness Evaluation

Soundness verification of utterance (a) boils down to proving the theorem:

$$P1: \quad (A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B) \vdash D$$

Analogously, for the backward reasoning step given in (d) we get:

$$P2: \quad E \vdash (D \vee E)$$

No specific requirements imposed on the proof system $\vdash$ (any kind of first-order theorem prover)

Soundness Evaluation (contd.)

Natural deduction calculus propably not appropriate: two intuitively very similar user proof steps may actually expand into natural deduction proofs of completely different size.

An important question concerns the appropriate choice of a proof system $\vdash$. Our hypothesis is that the abstract-level reasoning systems will provide more adequate measures for analyzing argumentative complexity of user proof steps since they better reflect human reasoning.

Empirical studies are possible and planned.

Granularity Evaluation

Requires analyzing the ‘complexity’ or ‘size’ of proofs: For utterances (a) and (d) above, it thus boils down to judging about the complexity of the proof tasks (P1) and (P2).

For illustration consider Gentzen’s ND as proof system $\vdash$.

Define *argumentative complexity*: number of $\vdash$-steps in the smallest $\vdash$-proof.

Judgements:

- argumentative complexity of (a) is bigger than that of (b)
- argumentative complexity of (a) is above a (tutorially) motivated complexity threshold

Relevance Evaluation

Alternative approaches listed according to increasing difficulty:

- Statically choose one or a few “golden proofs” and match the uttered partial proofs against them.
- Generate from the initially chosen golden proofs larger sets modulo, for instance, (allowed) re-orderings of proof steps and match against this extended set.
- Dynamically support relevance analysis with domain reasoning. For this, we test whether a proof can still be obtained from the new proof situation (using an abstract-level proof system). Resource-bound enumeration of possible proofs and proof step matching is additionally required.
- Stimulate research in proof theory: compact and tractable representation of the proofs in the proof space.

Relevance Evaluation (contd.)

Backward reasoning case (c): $D \vee E$ is refined to goal D.

Relevance question: can a proof still be generated? The task is thus identical to proof task (P1) as before.
A backward proof step that is not relevant according to this criterion is (d) since it reduces to:

$$P3: \quad (A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B) \vdash E$$

for which no proof can be generated. Thus, (d) is a sound refinement step that is not relevant, in contrast to utterance (c).

Approach needs to be refined: exclude detours and take tutorial aspects into account (teaching particular styles of proofs, particular proof methods, etc.).

Relevance Evaluation (contd.)

More challenging forward reasoning case is discussed next.
Example (a):

$$P4: \quad (A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B), D \vdash (D \vee E)$$

The question whether D is relevant reduces to the question whether there exists a proof for the given task that employs D and which is shorter than the best proof that can be obtained when deleting D from the available knowledge. According to this approach, utterance (b) describes an non-relevant proof step.

Note that we do not just ask about the existence of an arbitrary proof but about the existence of a proof with particular properties. This requires techniques such as (resource-bound and heuristic guided) enumeration of proofs.

What are the right provers?

For relevance and underspecification (see WP3), we need an approach that can (resource-bound and heuristically guided) enumerate at least some of the proofs in the proof space. For similar reasons as above, we assume that this mechanism should ideally operate on an abstract-level. Therefore, agent-based assertion level reasoning will be our first choice. As for granularity, we will investigate this hypothesis within small empirical experiments.