A Task/Entity-Based Context Model for Answering Follow-up Questions

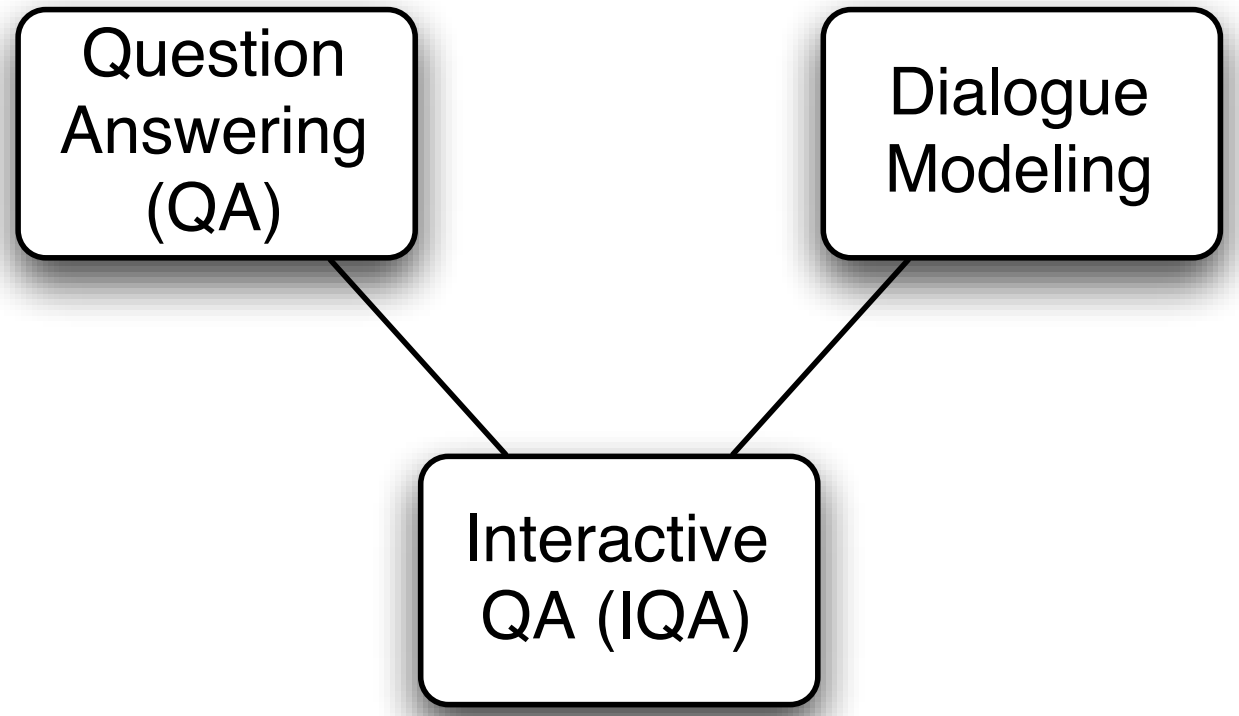
Raffaella Bernardi

co-worker: Manuel Kirschner

KRDB Center
Faculty of Computer Science
Free University of Bozen-Bolzano

April 30, 2008

Interactive Question Answering (IQA): Introduction



- QA systems retrieve answers to questions asked in isolation.
- DM tries to explain how human-human dialogue is structured.
- IQA extends QA systems with simple interactions (hence, questions are in a context) by taking advantage of what DM offers.

Context Modelling for IQA

- Main new challenge: Need of Context Modelling.
- IQA is similar to "Information Seeking Dialogues".

[Jönsson '97], [Stede & Schlangen '04] for Information Seeking Dialogues there is no need to:

- identify user's intention. By default: retrieve info.
- plan the dialogue. By default: Question-Answers pairs

On the other hand, there is the need of

- handling fragments.
- handling anaphora.
- being aware of the *local context*.

Problem: Questions within a Context (FUs)

User: How do I search for books?

BoB: You could search by "keywords", "title" and "author" by using the advanced interface.

User: Where can I find it?

BoB: At the url: `http://www.unibz.it/opacadvanced`

User: What about journals?

User: Can I do a guided tour?

BoB: The Library regularly offers guided tour.

User: OK, and when?

We will refer to these questions as Follow Up questions (FUs)

Coherence in Dialogues

- To develop a robust dialogue system able to process FU Qs and retrieve the appropriate answer, it is useful to be able to predict what a user is likely to ask about next.
- To meet this request, we assume that the user generally strives to engage in a coherent dialogue with the system.
- Thus, we have to answer the question how a coherent dialogue is structured.
- Coherence can often be defined in terms of *focus* and *focus shifts*.

1. which are the more plausible/natural/coherent transitions?
2. i.e., which is the most natural "focus flow"?

Focus in Dialogues

[Chai & Jin '04] "Context Question Answering":
"Focus": set of (related) "things" about which the user is seeking information.
The focus flows from one question to the next one, that could be about

- a different aspect of the focus ("topic exploration")
- a related focus (focus "coherent shift")

Overview on the notion of "focus" in User System Dialogues [Lecœuche et ali. 1998].

Research Hypothesis and Method

Hypothesis: Users have some specific library-related *task* (action, e.g. "search") in mind that they want to ask the system about, or they want to have info on some specific *entities* (e.g., "guided tour"). User will zoom into either of the two and might jump from the task (action/verb) to the entities that are possible fillers of the verb's argument slots.

Proposal We propose a task-and-entity-based view to describe the focus of questions and answers.

Focus in Dialogues

The IQA should be based on the following points:

1 Is the user interested in knowing something about a task?

1a If yes, is he asking a general question about this action or is he asking some specific question regarding some specific entities related to the task? (e.g. How to *search* for books vs. How to *search* for journals? vs. How to *search* via the advanced interface?)

1b If no (is not interested in a task), is he asking a question about an entity (e.g. which are the *opening hours*)?

We have identified the relevant tasks (search, reserve, read, etc.), and use PropBank as a base to define for each task the possible arguments, and assigned entities to the latter.

Examples: task and entity

1. User: How do I *search* for books?
2. BoB: You could search by "keywords", "title" and "author" by using the advanced interface.
3. User: Where can I find it?
4. BoB: At the url: `http://www.unibz.it/opacadvanced`
5. User: What about journals?

1. User: Can I do a *guided tour*?
2. BoB: The Library regularly offers guided tour.
3. User: OK, and when?

Empirical Data and Statistical Evaluation

We propose to:

1. Collect corpora of information seeking chats.
2. Observe how the focus flow within these dialogues.
3. Use these observations as base for our Information Seeking Chatterbot.
4. Evaluate it against the collected corpora.
5. Evaluate it against real users.

I will report about what we have done so far on points: 1-4.

Collection of Empirical Data

1. WoZ
2. Controlled Tests
3. Free FUs
4. On-line Chaterbot, BoB

Empirical Data: WoZ

a) 64 dialogues, with an average of 6 turns per dialogue.
b) Few switches to unrelated topics within the same dialogue.
c) Many follow up questions (FUs) to zoom into the topic (or a related one) or into entities introduced in the previous part of the dialogue.
d) 18,75% discourse phenomena (anaphoric pronoun, elided NP, fragments).
e) DP related to previous question or previous system response.

Hence, need to

- study the most probable "flow of focus".
- identify the size of the context (dialogue history) necessary to predict FUs.

Given e), a good description of the local context will help handling DP too.

Controlled Tests: Setting

Controlled FUs:

- 7 users,
- given 21 Dialogues (of different length),
- for each Dialogue, they were given one (or two) FUs on which the user had to give his/her preference (marks from 1 to 5).

Simplified Dialogues:

- User asks questions, the system gives answers
- Dialogs start with User Question and end with User Question.
- Only the user questions introduce new tasks.
- The system answers are chosen such that they concern exactly the tasks introduced by the corresponding questions, i.e., they do not introduce new tasks by themselves.
- Entities are introduced either by the system or the user.

Controlled Tests: example

31-QA5 (A5 vs. B5): CFL

1) **You:** How do I search for numbers?
2) **BoB:** In the advanced search there is an extra search field for ISBN and ISSN numbers.
3) **You:** Can I borrow media that do not have an ISBN nor ISSN number?
4) **BoB:** Yes, for example you can borrow DVDs, too.

* **1A5: 5A) You: Can I borrow journals?**

(M1, D0; TS1, CFL; TS FU: no.)

Please choose **only one** of the following:
- ☐ Very unnatural
- ☐ Unnatural
- ☐ Neither natural nor unnatural
- ☐ Natural
- ☐ Very natural

* **1B5: 5B) You: How do I search for DVDs?**

Memory 2, Distance 1; TS2, CFL; TS FU: yes.

Please choose **only one** of the following:
- ☐ Very unnatural
- ☐ Unnatural
- ☐ Neither natural nor unnatural
- ☐ Natural
- ☐ Very natural

A5) 1 (search) $\rightarrow$ 2 (borrow) $\rightarrow$ 2 (borrow) vs. B5) 1 (s.) $\rightarrow$ 2 (b.) $\rightarrow$ 1 (search).

Controlled Tests: purpose

topic exploration a different aspect of the focus

1. asking info about the same task and same argument (e.g. search/books: where, when)
2. asking info about the same task but different argument (e.g. search: books, journals)
3. asking more info about the same entity (guided tour: when, where)

coherent shift a related focus

1. going from a node in the task-structure to a related node of the entity-structure and vice versa.
2. going from a task to a related one. (not done)
3. going from an entity to a related one. (not done)

shift an unrelated focus.

Controlled Tests: results

Relevant differences:

Distance It's worth for an IQA to remember the previous focus; it's not worth adding further memory.

Entity there is a preference for FUs whose focus is on the same Entity which was the focus of the previous question.

Task, Entity, Question The preferred ones is: (1.) Same Task, same ARG/ENTITY, Different Q-Type; (2.) Different TASK, same ENTITY (same ARG?)

Empirical Data: Summary of analysis

- WoZ and Controlled Test: useful to get a first approximated answer
- BUT: many variables enter into the picture and are difficult to control. Difficulties in setting up the dialogues.
- BUT: users feel controlled and this might influence their answers.
- BUT: nr. of dialogues limited –to be checked manually.

Good starting point for our next experiment based on Regression Models.
These first tests gave us an idea of which features might play a role in "Context Modelling" for IQA.

Relevant Features

- 3 surface-based features: which {Task, Entity, QuestionType} are identified in the user question
- 1 task structure-based feature: how many of the candidate task's participant entities (as encoded in the task ProbBank like frame) are identified in the user question
- 4 focus continuity features: whether Task, Entity or QType are continued in the user question, wrt. previous dialogue.
  - Task, Entity, QType continuity wrt. previous user question
  - Entity continuity wrt. previous system answer
- 2 task structure + focus continuity features:
  - Focused Task of previous user question has candidate entity as a participant
  - Task candidate has focused Entity of previous question as a participant

Prototype

Have built prototype IQA system incorporating new Q→A mapping algorithm

- based on analyzing user Q via <Task, Entity, QuestionType>
- context-dependent FU Qs are "completed" with information from dialogue history (i.e., focused things) (cf. e.g., RITEL system)
- system provides a testbed for systematically experimenting with all parameters of the Q→A mapping algorithm

The Q→A Mapping Algorithm

- We hand-assigned <Task, Entity, QuestionType> to each answer of our repository (> 200). Let A be our repository.
- For each new user question q, score each answer $a \in A$ based on the values of the k features $x_{1,q,a} \ldots x_{k,q,a}$ (in our case $k = 10$)
- return the highest-scoring answer $\hat{a}$

$$\hat{a} = \text{argmax}_{a \in A}(\beta_1 x_{1,q,a} + \cdots + \beta_k x_{k,q,a})$$

Empirical Data: annotated answers

`Task: Search` has the following arguments:

- ARG-0 (e.g., user),
- ARG-1 (e.g. book),
- ARG-2 (e.g., search terms),
- ARG-LOC (e.g. library location),
- ARG-MNR (eg. advanced interface)

The Answer: "You can restrict your query in the OPAC on individual Library locations. The search will then be restricted e.g. to the Library of Bressanone-Brixen or the library of the MUSEION. Do you want to know how it works?"

is marked by: `Task: Search, ARG-LOC: Library Location, Qtype: Yes-No`

Feature Scores ($\beta_1, \ldots, \beta_{10}$)

To each of our 10 features we chose an intuitive score. Each β potentially contributes to total score of each answer candidate *a*.

- surface-based features:
  - (0 vs. 4) user question Q type equals Q type assigned to the answer in the focus structure
  - (0 vs. 3) user Q contains the stem of the task assigned to the answer
  - (0 vs. 2*n) user Q contains the stem of the entity assigned to the answer
- task structure-based features:
  - (0 vs. 1*n) user's question contain entities that are filler of task's arguments.
- Example focus continuity-based features:
  - (0 vs. 1) task continuity
  - (0 vs. 1) entity continuity

Prototype: Example

- First Q: "Can I get search results for a specific library location?"
  (search, argm-loc, librari locat, yesno_TASK)
- FU: "In case the book I am looking for is only available in Brixen, can I ask to get it delivered to Bolzano?"
- gets the correct answer which is tagged as:
  Task-Entity Question: pick up, arg2, bolzano, yesno-TASK
- Although the user did not use "pick up", the algorithm identified the correct task with the help of the number of matched entities.
- Here are the scores our algorithm got for this answer, which sum up to 10, making it the top-scoring answer:
  qTypePatternMatches=0, qTypeEqualsTopOfStack=1, taskStemMatches=0, taskEqualsTopOfStack=0, entityStemMatches=5, entityEqualsTopOfStack=0, *matchedEntitiesInTask=2*, participantSlotEqualsTopOfStack=0. (8)

Empirical Data: Gold standard

We need test data (a "gold standard" annotation) to run the Q→A algorithm on.

1. collect Q-A-FU triple.
2. have them annotated with <Task, Entity, QuestionType>

Empirical Data: Corpus

- Introduction to part 1

Introduction to Part 1

On the following page we will show you 13 short human-computer dialogues, each consisting of a first question by the user, and a corresponding answer by BoB.

Your task:

Take the role of the user, and imagine you were really using BoB to get information about different library topics. For each of the 13 dialogues, provide **a follow-up question that will help further serve your information need in that specific situation.**

0- FUFU

* **01: 1) You: What services does the library offer?**
2) BoB: In addition to a substantial collection of scientific literature, the university library provides information and advisory services, an interesting selection of training courses, a document delivery service and much more.
3) You:

Please write your answer here:

8 users, 11 Dialogues.

Evaluation of the Q→A Algorithm

- Evaluation of the Q→A algorithm in terms of answer accuracy (%)
- We do not consider the answers themselves, but the "key" <Task, Entity, QuestionType>
- Idea: if the key triple is "correct", the corresponding answer (the canned-text answer stored under this triple) presumably is so, too

Accuracy results from experiments: 24 out of 78 questions answered correctly (30.8% accuracy)

Next Steps

1. The annotated corpus has been used as training data to learn how "relevant" each of the features we have identified are in order to detect an answer in an IQA setting.
2. Inter-annotator Agreement (objective measures)
3. BoB on-line and collect real users' dialogues. (June?)

I will now report on 1.

Learning Feature Coefficients with a Logistic Regression Model

- Idea: have a more principled way of setting the "scores" that each of our 10 features contributes to the total score for $\hat{a}$
- Logistic regression models (e.g. Agresti 2002) describe the relationship between some predictors (i.e., our features) and an outcome (answer correctness)
- We use the logit beta coefficients $\beta_1, \ldots, \beta_k$ that a logistic regression model estimates for each predictor (from training data, using maximum likelihood estimation) as our empirically learned scores
- Nice aspect: yields human-readable learned coefficients, showing contribution of each predictor

Generating Training Data from Human Annotations

- Again, use our "gold standard" annotations we've used for testing (but split training/test data)
- For each human-annotated question q and each answer key triple from our repository ($a \in A$), calculate the values for the k=10 features

$$x_{1,q,a}, \ldots, x_{k,q,a}$$

- if the key triples <Task, Entity, QuestionType> of q and a are identical, take the feature vector as a training instance for a correct answer
- else: take the feature vector as a training instance for a wrong answer (we get $|A| - 1$ false for each correct instance, seems to work)

Learned Coefficients

Example learned model with only the 6 significant factors (continuous factors standardized as z scores for comparability):

$$\text{Prob}\{\text{answerCorrectness} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

$$\begin{aligned} X\hat{\beta} = \; & -7.9760 + 2.5396\ \text{qTypePatternMatches} \\ & +2.1650\ \text{taskEqualsTopOfStack} \\ & +1.3683\ \text{taskStemMatches} \\ & +0.4624\ \text{lengthOfEntityStemMatches} \\ & +0.3787\ \text{entityIsFromSystemAnswer} \\ & -1.2427\ \text{taskTakesEntityOnStack} \end{aligned}$$

Evaluating Answer Accuracy with Learned Coefficients

Plugging the learned coefficients $\beta_1, \ldots, \beta_k$ into the $Q \rightarrow A$ algorithm, and testing unknown data (i.e., *different person's* annotation of questions):

$$\hat{a} = \text{argmax}_{a \in A}(\beta_1 x_{1,q,a} + \cdots + \beta_k x_{k,q,a})$$

Accuracy results from experiments:

- (Intuitive scores from previous slide: 24 out of 78 questions answered correctly (30.8% accuracy))
- learned coefficients (only 3 surface features in model): 40 out of 78 questions answered correctly (51.3% accuracy)
- learned coefficients (all 6 significant features): 47 out of 78 questions answered correctly (60.4% accuracy)

Projects

- QType detection (Dinh Le Thanh –EM in LCT student)
- Controlled Natural Language and Chatterbot
- Shallow reasoning: e.g. question about "other xx"
- Reasoning module (started with (Marija Slavkovik, –EMCL student)
- Extraction of focus structure from a collection of documents.
- User evaluations
- Evaluation: BoB (Regular Expression) vs. BoB (Context Modelling)
- Addition of more elaborate NLP tools (PoS tagging of question, parsing, Lexical Resources etc.)
- Multilingual evaluation (comparison of German, Italian and English Dialogues).
- ...