

# From artificial questions to real user interaction logs: Real challenges for Interactive Question Answering systems

Raffaella Bernardi and Manuel Kirschner

KRDB, Faculty of Computer Science  
Free University of Bozen-Bolzano, Italy

## Abstract

Much research in Interactive Question Answering (IQA) has centered on artificially collected series of context questions. Instead, the goal of this paper is to emphasize the importance of evaluating IQA systems against *realistic* user questions. We do this by comparing the highly popular TREC QA context task data against two more realistic data sets: firstly, a corpus of real user interaction logs that we collected through a publicly accessible chat-bot, and secondly, a corpus of QA dialogues collected in a Wizard-of-Oz study. We compare these data using basic quantitative measures and different measures for expressing inter-utterance coherence. We conclude with proposals for choosing test data for a new evaluation campaign that is centered on realistic user-system interactions, and that is well suited for empirical and Machine Learning approaches.

## 1. Introduction

Question Answering (QA) systems have reached a high level of performance within the *single, factoid question* scenario originally defined by the TREC QA competitions. As a consequence, the research community has moved on to tackle new challenges, as shown by the *context question* and *Interactive QA (IQA)* tasks proposed in recent instantiations (Voorhees, 2004). The idea of extending “single shot” questions to context QA first made its way into the 2001 QA track (Voorhees, 2001), in the so-called context task, which consisted of 10 question series of around 5 topically related questions each. From TREC 2004 (Voorhees, 2004) onward, the *main* QA task was changed to consist of series of questions, where for each series a so-called *target* string explicitly identified the topic that is common to all the questions from a particular series. Examples of TREC’01 and TREC’04 series of questions are shown below.

1. In what country was the first telescope to use adaptive optics used?
2. Where in the country is it located?
3. What is the name of the large observatory located there?

Table 1: Sample from TREC’01

- Series question target:* Hale Bopp comet
1. When was the comet discovered?
  2. How often does it approach the earth?
  3. In what countries was the comet visible on its last return?

Table 2: Sample from TREC’04

The yearly TREC QA track – run by NIST from 1999 through 2007 (e.g., (Dang et al., 2007)) – has helped the QA community share results, and lead to new techniques being embraced much faster.

In a report known as the ARDA QA Roadmap (Burger et al., 2000), a number of researchers from the QA community suggested several important research challenges for QA. Two of the challenges mentioned were Context QA IQA, both of which were later addressed in specific tracks of the TREC QA task, and thus have had a major influence on current QA research in general.

For the case of Context QA, (Burger et al., 2000) see the role of context as that of clarifying a Follow-Up Question (FU Q), resolving ambiguities, or keeping track of an investigation performed through a series of questions. The underlying motivation is that in a real information-seeking scenario, questions are not asked in isolation; instead, users ask FU Qs that might relate in different ways to the ongoing dialogue. In this work, we confirm this claim empirically by analyzing the corpus of real user interaction logs that we have collected through BoB, a chat-bot that has been providing library help-desk information on our University Library web-site<sup>1</sup> for over one year.

As for IQA, (Burger et al., 2000) foresee that a questioner might want not only to reformulate a question, but to engage in a real user-system dialogue; thus, IQA tries to go a step beyond Context QA, towards a truly intelligent system for humans trying to access information in an efficient way. In practice, the term *IQA* has been used to denote quite different extents of dialogue capabilities of a QA system. One example is the TREC complex Interactive QA (cIQA) task, conducted as part of the 2006 TREC QA track (Kelly and Lin, 2007): in an approach inherited from interactive IR, assessors had to provide one iteration of user-system feedback in the form of relevance feedback; the system was then supposed to take this feedback into account and provide a new and improved answer to the user. Besides cIQA, there has been considerable work on IQA systems with even more extensive interactive capabilities. Such systems were characterized in the following ways by contributors of a recent, influential IQA Workshop (Interactive Question Answering Workshop at HLT-NAACL, 2006): the system draws the user into a conversation (Strzalkowski, 2006); the

<sup>1</sup><http://www.unibz.it/en/library>

system understands what the user is looking for, what the user has done and what the user knows (Kelly et al., 2006); the system is a partner in research. Other recent approaches considering highly interactive IQA systems can be found in (Maybury, 2003), (Strzalkowski and Harabagiu, 2006) and (Mitkov et al., 2009).

In this work, we propose a definition of *IQA* that includes *Context QA* (as in the TREC QA context track), but puts emphasis on the availability of *realistic* user questions, and on the existence of system answers in the dialogue context. We believe it is crucial that a large enough amount of IQA dialogue data are easily available for empirical and Machine Learning-based research in IQA: this criterion was met in the case of the TREC Context QA data, but not in the case of TREC cIQA data, nor of much of the above-mentioned literature proposing more sophisticated, highly interactive IQA systems. The BoB dialogue data described in this paper try to strike a balance between the two goals: the availability of IQA data for empirical research, combined with an adequate level of realism and naturalness of the IQA dialogues.

The goal of this paper is to emphasize the importance of evaluating (Interactive) QA systems against *realistic* user questions. In order to highlight aspects of real user interactions with an IQA system, we compare the TREC context task data against two data sets: firstly, a corpus of QA dialogues collected semi-artificially in a Wizard-of-Oz study (Bertomeu, 2008), and secondly, BoB’s real user interaction logs (Kirschner, 2010).

We compare the data sets by considering basic quantitative measures like dialogue and utterance length, prevalence of anaphoric references, and different measures for quantifying inter-utterance coherence. As for the latter, we consider those measures that we have used successfully as features for modeling IQA dialogue structure, and that improved an IQA system’s accuracy in answering FU Qs (Kirschner et al., 2009; Bernardi et al., 2010). In particular, we use a *shallow* feature that quantifies inter-utterance coherence through simple string similarity, and a variety of *deep* features that define coherence based on two existing theories of coherence in Dialogue and Discourse. The first theory (Sun and Chai, 2007), strongly related to the well-known Centering Theory (Grosz et al., 1995), looks at *entity*-based coherence, while the second theory (Chai and Jin, 2004) considers *action*-based coherence, where the verbs of successive user questions are considered.

We now move to describe the different IQA data sets (Section 2.), introduce the inter-utterance coherence features (Section 3.), and the data sets in terms of these features (Section 4.). From this analysis, in Section 5., we draw conclusions that could be useful for setting up a new evaluation campaign for Interactive Question Answering systems that considers *realistic* user questions.

## 2. Data sets

We now introduce each of the data sets, explaining how it was collected, and providing a qualitative description on the linguistic and stylistic levels. We also provide an example excerpt for each data set. Basic quantitative features of the data sets will be given in Section 4.1..

**TREC** The TREC data come from the Text REtrieval Conferences question answering (QA) track (Voorhees, 2004), namely from its context task. This task was designed to study contextual, interactive QA by allowing for series of contextually related questions. We use two English language data sets from the 2001 and 2004 editions of the TREC QA track.

As the two data samples in Tables 1 and 2 show, all questions are grammatical sentences and contain no typos, in stark contrast to any real user scenario. Moreover, again differently from a real setting, Follow-up Questions were asked by the NIST staff without knowing the answer to the previous question. Finally, within a series there are no topic shifts, and the FU Qs are always about the same topic (question target) defined for the entire series. All these aspects show how TREC data are rather different from real user-system interactions.

**Bertomeu** The corpus described in (Bertomeu, 2008) was collected via a Wizard-of-Oz study, involving 33 subjects. Interactions are about language technology, and the information comes from a language technology information database. The subjects were given the following tasks. Through their interactions, they had to find (i) three traineeships at three different projects in three different institutions, (ii) three conferences in the winter term and three conferences in the summer term taking place at different times, on different topics, etc., (iii) information for writing a report on language technology in Europe in the last ten years. Subjects were asked to complete these tasks in one hour. The wizard answered mostly by listing entities from a database, or by asking a clarification question and making statements about the database contents. Table 3 shows an excerpt from a user-wizard dialogue.

Q1: Can you show me those which are about dialog processing or artificial intelligence?
A1: The server is busy. Your question will be processed in a moment . . . International Joint Conference on Natural Language Processing.
Q2: Are all these conferences taking place between now and September 2006?
A2: yes
Q3: When does “Translating and the computer 27” take place?
A3: From 2005-11-24 until 2005-11-25.
Q4: I’m interested in “International Conference on Image Processing” and “31st International Conference on Acoustics, Speech, and Signal Processing” and “13th Annual ACM International Conference on Multimedia”. When do these take place?
A4: From 2005-11-11 until 2005-11-15; From 2006-05-14 until 2006-05-19; From 2005-11-06 until 2005-11-12.
Q5: When is the “Second Australian Conference on Artificial Life”?
A5: From 2005-12-05 until 2005-12-08.

Table 3: Sample from Bertomeu

Differently from TREC data, the interactions with the wizard contain ungrammatical sentences as well as sentences with typos; moreover, the users asked FU Qs after hav-

ing received an answer. These features make the Bertomeu data more related to real user-system interactions. However, since the subjects were given precise tasks and a time limit to achieve their information goals, the data is inherently not natural.

**BoB** This data set was collected via the Bolzano Bot (BoB), which has been working as an on-line virtual help desk for the users of our University Library since October 2008.<sup>2</sup> Bob speaks English, Italian and German; in this work, we consider only the English logs. Users ask questions and the system answers them – apart from a few sub-dialogues, tagged as such, in which BoB asks clarification questions or suggests other relevant information the user might be interested in.

<p>Q1: how can i request a book that is not in the library  A1: If we don't have the book you should in any case check the catalogue of the state libraries Claudia Augusta or Tessmann. If it isn't there as well you can order it via our Interlibrary Loan service.  Q2: explain the interlibrary loan service  A2: If you need a book or an article which is not available in any library in South Tyrol you can use our Interlibrary Loan Service. Would you like to know more about it?  Q3: yes  A3: What exactly do you need? Some general information about the Interlibrary Loan Service? Or do you want to know how to order a book? Or do you need an article?  Q4: how to order a book  A4: To order a book we need the following data: the author of the book, the title of the book, and the publication date.</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 4: Sample from BoB

Like Bertomeu's logs, the interactions with BoB contain many typos and ungrammatical sentences, as well as just keywords. Moreover, the dialogues are rather short in length: while some users seem to use the system to explore library-related information and let themselves be "guided" by BoB, many users seem to have just one information need, and leave the conversation after they asked the relevant question, and hopefully received the correct answer. Moreover, there are several FU Qs that are paraphrases of previous questions. In this case, the user might be trying to refine her question, because the answer was correct but not what the user wanted to know, or the answer was incorrect and the user thinks the system has not understood her question. Another possible cause for rephrased questions is that the user explores the topic further by moving the focus of attention to a new related entity or a new related action, as in the following example:  $Q_1$ : *Could you recommend me some book?*  $Q_2$ : *Could you recommend me some novel?* These kinds of interactions seem typical of real user data, and they are also reported in the literature (Bertomeu, 2008; Yang et al., 2006). Like in TREC data, FU Qs that are *Topic*

<sup>2</sup>We developed the chat-bot web application as an open source project, which we would like to share with the research community interested in collecting similar IQA dialogues. See <http://code.google.com/p/chatbot-bob>.

*Continuations* – i.e., that do not switch to some unrelated new topic – may contain ellipses and anaphora, as in:  $Q_1$ : *Where can I find design books?*  $Q_2$ : *and dvd?*. We will address this aspect in Section 4..

### 3. Inter-utterance features

It has been shown that an IQA system generally needs to consider just the immediately preceding interactions (i.e., the previous question and its answer) to answer FU Qs (Kirschner et al., 2009). We converted all IQA dialogues of the data sets described above into what we call *dialogue snippets*, each consisting of four successive utterances:  $Q_1$ ,  $A_1$ ,  $Q_2$ ,  $A_2$ . Each snippet thus represents a FU Q, termed  $Q_2$ , preceded by the previous user question and system answer, and followed by its (correct) answer  $A_2$ .<sup>3</sup> We use this snippet representation to calculate the two types of inter-utterance features described in the following.

#### 3.1. Shallow string similarity feature

We use a shallow feature to measure string similarity between two utterances within a snippet. The idea is that string similarity is a simple approach to measuring *coherence* between two utterances; we want to compare coherence between  $Q_2$  and the preceding utterances across the different IQA dialogue sets to get a first, shallow approximation of how FU Qs relate to the dialogue context. Our string similarity metric is based on inverse *tf.idf*-distance of the bag of words representations of the two utterances. If two utterances (e.g.,  $Q_1$  and  $Q_2$ ) share some terms, they are similar; the more *discriminative* the terms they share, the more similar the utterances. See (Kirschner, 2010) for a detailed technical description of our implementation of this feature.

#### 3.2. Dialogue and Discourse features

Like the shallow feature introduced above, the Dialogue and Discourse features described in this section measure different types of coherence between utterance within a dialogue snippet. However, now the notion of coherence is based on different theories from the field of Dialogue and Discourse modeling. Following (Sun and Chai, 2007), we consider features describing coherence in terms of repeated occurrences of discourse entities: entity-based coherence. Moreover, following (Chai and Jin, 2004), we define features that describe different *Informational Transitions* holding between a user's previous question and their FU Q, based on the actions (i.e., the verbs) underlying these questions.

##### 3.2.1. Entity-based dialogue coherence

We introduce three features for describing coherence relations between specific pairs of utterances, based on the reference, forward and transition models of (Sun and Chai, 2007). These relations define dialogue coherence by checking for the repetition of certain discourse entities, i.e., noun phrases, within a dialogue snippet. The three relations are inspired by Centering Theory (Brennan et al., 1987; Grosz

<sup>3</sup>Because some of the features described below need to consider also the preceding context of  $Q_1$ , we keep information about the order in which the snippets represent the original dialogue.

et al., 1995); more specifically, their definitions build on the following definitions from (Brennan et al., 1987):

*Forward-looking centers*: each utterance is associated with a list of *forward-looking centers*, consisting of those discourse entities that are mentioned in the utterance.

*Preferred center*: the list of forward-looking centers is ordered by likelihood of each entity to be the primary focus of the subsequent utterance; the first entity on this list is the *preferred center*.

Our implementation of the three Centering-Theory-based features relies on the automatic detection of forward-looking and preferred centers, and on automatic anaphora resolution. For these tasks, we make use of GuiTAR (Poesio and Kabadjov, 2004; Kabadjov, 2007). Firstly, GuiTAR yields a list of resolved antecedents referred to in a given utterance by anaphora.<sup>4</sup> Secondly, it finds a list of an utterance’s forward-looking centers, i.e., any noun phrase directly mentioned in the utterance. In this work, and following (Ratkovic, 2009), we consider the *preferred center* to be that entity from the list of forward-looking centers which is *mentioned first* in the utterance, and which is not a first or second person pronoun.

We use the following approach, proposed in (Ratkovic, 2009), to identify the *preferred center* of each question. For all anaphora found in the question, we use GuiTAR to extract their antecedents, again using the previous questions as context; the first (in terms of linear order) antecedent which is not a first or second person pronoun<sup>5</sup> becomes the *preferred center* of the question. If no preferred center was found so far, the first noun phrase (which is not a first or second person pronoun) appearing in the question itself becomes the preferred center.

Our first feature, `center.Reference`, implements the idea behind the *reference model* of (Sun and Chai, 2007). It is a binary feature that indicates whether a specific coherence relation holds between  $Q_2$  and  $A_2$ . First of all, we resolve any anaphora present in  $Q_2$ , providing  $Q_1$  as dialogue context. Note that the dialogue context does not include the preceding answers. Although we show in Section 4.2. that these answers are likely locations of antecedents to anaphora found in FU Qs, we do not consider answers in the feature definition for purely practical reasons (discussed in Section 4.2.), to keep our data sets comparable. The `center.Reference` feature evaluates to *true* if the noun phrase head of any antecedent is mentioned in  $A_2$ . Note that in our implementation we do not consider cases that are string-identical, thus disregarding all classes of anaphora detected by GuiTAR, but personal pronouns.

The `center.Forward` feature implements the *forward model* of (Sun and Chai, 2007). It is again a binary feature, this time indicating the presence of a specific coherence relation holding between  $Q_1$  and  $A_2$ . After resolving

<sup>4</sup>Anaphora considered by GuiTAR are: definite noun phrases, proper nouns, proper nouns with definite articles, and personal pronouns.

<sup>5</sup>Very often in IQA dialogue data the subject of the question is a personal pronoun like “I”. This pronoun carries no useful information regarding the informational content of the question, and we thus exclude such pronouns from our algorithm.

anaphora in  $Q_1$  – using  $Q_2$  from the previous dialogue snippet as context – the `center.Forward` feature becomes *true* if either the noun phrase head of any antecedent is mentioned in  $A_2$ , or any *forward-looking center* from  $Q_1$  can be found also in  $A_2$ .

Finally, the `center.Transition` feature is based on the *transition model* of (Sun and Chai, 2007). It builds on the four discourse transitions between adjacent utterances that Centering Theory introduced (Brennan et al., 1987). Somewhat differently from that classic theory, (Sun and Chai, 2007) define the transitions depending on whether the head and/or the modifier of the *preferred centers* are continued or switched between  $Q_1$  and  $Q_2$ .<sup>6</sup> The four possible values of the `center.Transition` feature are defined as follows, based on the two preferred centers of  $Q_1$  and  $Q_2$ : *Continue*: both the head and the modifier stay the same. *Retain*: the head stays the same, but the modifier is different. *Smooth shift*: the head is different, but the modifier stays the same. *Rough shift*: both the head and modifier are different.

### 3.2.2. Action-based dialogue coherence

We use three different features to describe the Informational Transitions proposed by (Chai and Jin, 2004). All these are based on certain relations between the predicate-argument structures of two consecutive user questions,  $Q_1$  and  $Q_2$ .

- (a) `ConstraintRefinement`: a question concerns a similar topic as that of a previous question, but with different or additional constraints
- (b) `ParticipantShift`: the FU Q is about a similar topic but with different participants
- (c) `TopicExploration`: the two questions are concerning the same topic, but with different focus

We implemented these features based on the grammatical relations produced by the Stanford parser (Klein and Manning, 2003) in dependency mode (de Marneffe et al., 2006). The main ideas behind the feature implementations are the following (see (Ratkovic, 2009; Bernardi et al., 2010) for more details):

- (a) the two questions contain the same syntactic predicate and the same subject or object, but  $Q_2$  has either an *additional* or a *missing* argument (subject, object, adverb, preposition, or adjectival modifier) when compared to  $Q_1$
- (b) the two questions have the same syntactic predicate, but either the subject, object or argument of some preposition are different
- (c) the two questions have either the same syntactic predicate, subject, object or preposition.<sup>7</sup>

<sup>6</sup>Centers are noun phrases. The syntactic structure of a noun phrase comprises a *head noun*, and possibly a *modifier*, e.g., an adjective.

<sup>7</sup>We found this rather lax definition to work best in FU Q classification experiments.

## 4. Data comparison

### 4.1. Basic quantitative measures

Differences among the dialogue data sets are already evident by looking at basic statistics underlying the data.

Table 5 provides important quantitative measures. Most evidently, due to the Wizard-of-Oz design, the dialogues in the Bertomeu data are significantly longer than the naturally occurring IQA dialogues of genuinely interested users in the BoB data. Also, the questions are twice as long on average, indicating that users tend to form simpler and shorter queries in an actual IQA system. For the BoB data, note that before extracting dialogue snippets from the 1,161 dialogues containing at least one FU Q, we removed those dialogues where  $Q_2$  is not a library-related question, i.e., where the user did not seem to have an information need.

Looking more in detail at the dialogue lengths in BoB, Table 6 gives the counts and proportions of dialogues with the typical numbers of user questions. In this realistic IQA setting, two thirds of users asked at least one FU Q. The mean number of user questions across all dialogues containing at least two user questions is 5.3.

### 4.2. Anaphora

To assess the relevance of the dialogue context preceding  $Q_2$ , we again resort to GuiTAR for detecting and resolving anaphora; we now compare the resulting anaphora counts across the different IQA data sets. Table 7 lists counts and corresponding proportions out of the total number of  $Q_2$ s of each data set. From this table, we note the following: both the TREC and Bertomeu data sets contain proportionally more total anaphora than the realistic IQA data from BoB. The difference in anaphora proportions between BoB and Bertomeu is mostly due to personal pronouns (pers-pro) and proper nouns (pn): in both categories, Bertomeu contains twice as many anaphora than BoB. On the other hand, both TREC data sets contain a clearly exaggerated proportion of personal pronouns, with respect to both BoB and even Bertomeu. This shows again that the TREC question series can not be taken to represent realistic IQA questions.

Although we show in Table 8 that previous answers are likely locations of antecedents to anaphora in questions, we do not provide GuiTAR with the previous answer  $A_1$  as context for purely practical reasons. Firstly, for TREC, answers are not available, and secondly, for Bertomeu, the syntactic parser used by GuiTAR fails on the majority of system answers from that data set, due to their excessive sentence length. However, for the BoB data set, Table 8 does explore the issue of considering  $A_1$  as additional context in the anaphora detection and resolution phase. From this table it is evident that the previous answer plays an important role as a location for antecedents: if GuiTAR considers also  $A_1$  as a potential location of antecedents for anaphora from  $Q_2$ , there is a relative increase of 61% of detected anaphora.

Finally, to get a rough estimate of the accuracy of our automatic anaphora detection procedure based on GuiTAR, Table 9 compares automatically detected anaphora against a gold standard hand annotation. We use the two TREC data sets for this purpose. From this table it seems evident

that GuiTAR has a problem with recall, i.e., it seems to miss anaphora that were found by the human annotator. We still believe that our automatic procedure serves its purpose as a means for automatically comparing the IQA dialogue sets we are interested in.

### 4.3. Inter-utterance features

**String similarity feature** We calculated the shallow, string similarity-based feature as described in Section 3. to express the degree of term-based similarity between two consecutive questions across the different data sets. TREC data contain only topic continuation (TC) FU Qs, whereas BoB logs contain topic shifts (TS) too. Hence, we took a sample of BoB's logs (417 snippets out of 1,522) and marked manually whether the FU Q was a TC or a TS; the sample snippets contain 250 TC and 167 TS FU Qs. In Table 10 we summarize the average of the similarity between a FU Q and its previous question. In the case of BoB, we report string similarity figures of both the whole dialogue corpus, as well as the subset containing only those 250 FU Qs marked as TC.

We make two observations based on this table. Firstly, across all data sets, the transitions between  $Q_1$  and  $Q_2$  have the highest average string similarity of all utterance-utterance combinations. This is a first indication that consecutive questions in IQA often concern similar topics, by way of containing similar terms. As we see from the higher similarity scores of  $Q_1.Q_2$  for the TC subset of the BoB data compared to the full BoB data, topic continuation seems to be detectable to some extent already with this simple shallow feature of string similarity. The second observation we draw from this table is that the average string similarity between  $Q_2$  and its correct answer ( $A_2$ ) is lower for the Bertomeu data set when compared to the BoB data. We attribute this difference to the inherently different nature of questions and answers across the two data sets; as shown in Section 2., BoB answers consist of highly grammatical English sentences, while Bertomeu questions and answers tend to consist of long lists of dates or proper names.

**Dialogue and Discourse features** Table 11 shows how the different dialogue data sets differ in terms of our Dialogue and Discourse features introduced in Section 3.2.. The percentages in the table represent the proportions of the data sets for which the respective features hold (i.e., evaluate to true, or to one of the four `center.Transition` values).

Regarding dialogue coherence in terms of the Centers, we make two observations from Table 11. Firstly, we note that compared to the realistic IQA data from the BoB data set, the Bertomeu data exhibit much lower counts of dialogue snippets where the `center.Forward` feature holds, i.e., where there is entity-based continuity between  $Q_1$  and  $A_2$ . We attribute this difference to the typically list-like structure of  $A_2$  in the Bertomeu data; the large difference in proportions indicates some unnatural property of the Bertomeu data. As for the Centering-Theory-based transitions between  $Q_1$  and  $Q_2$  described by the `center.Transition` feature, we note that the TREC data exhibit a rather large proportion of *continue* transitions; the numbers suggest a closer structural similarity of

	BoB	TREC'01	TREC'04	Bertomeu
Dialogues (= QA sessions)	1,161 <sup>a</sup>	10	64	33
Number $Q_2$ s (= nr. snippets)	1,522	32	221	1,052
Mean utterances per dialogue/QA session	3.86	4.2	4.47	66.2
Mean $Q$ length (words)	4.4	7.7	6.0	8.8
Mean $A$ length (words)	26	–	–	118

<sup>a</sup> 1,161 (or 66%) from a total of 1,765 dialogue sessions contained at least one FU Q.

Table 5: Quantitative measures of data sets

Nr. of user questions in dialogue	1	2	3	4	5	6	>6
Nr. of dialogues (tot.: 1,765)	604	313	246	160	112	73	257
Proportion of tot. dialogues	34.2%	17.7%	13.9%	9.1%	6.3%	4.1%	14.6%

Table 6: Counts and proportions of BoB dialogues along numbers of user questions

the TREC data to the topic continuation (TC) subset of the BoB data. This supports our observation that the questions in TREC cannot be taken to represent real user questions in IQA dialogues.

Finally, looking at the action-based coherence features in the last three rows of Table 11, the Bertomeu data show similar feature proportions as the BoB Topic Continuation (TC) data set, but rather unlike the complete BoB data set. This is another sign that the Bertomeu data do not represent naturally occurring topic shifting behavior between user questions, but rather seem to exhibit Topic Continuation properties.

## 5. Conclusion

We believe the IQA research community could benefit from a new evaluation campaign in the style of the TREC QA context track, but resolving its two shortcomings: the artificiality of user questions, and the lack of preceding system answers on which FU Qs might build. We would hope and expect such a campaign to give rise to a new wave of research in the area of discourse and context modeling, which would aim to improve an IQA system’s ability to answer user FU Qs. In this paper, we have introduced and described our collection of BoB IQA dialogue data, and have shown how these data compare to other relevant data sets.

In Section 3. we introduced several measures for quantifying relations between utterances in IQA dialogue snippets, based on either string similarity, or different theories of Dialogue and Discourse coherence. We claim that these methods provide important insights into the inter-utterance structure of IQA dialogue data, and allowed us to point out relevant differences between the realistic BoB data set and two less natural data collections. The goal for a new evaluation campaign should be to provide a large set of IQA dialogues that resemble realistic data in different aspects, such as the features and measures we have introduced in this paper. We have introduced the set of BoB IQA dialogues as our attempt to provide such a data set to the research community.

In Section 4. we used the above-mentioned inter-utterance measures to pinpoint relevant differences between the data sets. However, we started by exploring differences that became evident already through the comparison of some

quantitative measures. First of all, real users in an IQA setting do ask FU Qs: in the case of BoB, two thirds of all IQA sessions contained at least two user questions. As opposed to artificially collected user questions, real user questions tend to be relatively simple and short, with an average word length of 4.4 words. Comparing this average to the same measure calculated from the query logs of a commercial web search engine, which was 2.35 in the year 1998 (Silverstein et al., 1998), we note that realistic IQA user questions fare somewhere in between web search engine queries and questions from the more artificial IQA data sets we have analyzed here. Interesting steps for further research would be to analyze and compare web search query logs using the different measures that we proposed here, and to see if over the course of the last decade web search queries might have evolved towards longer, and maybe also more contextually related queries.

From our analysis of real IQA interaction logs regarding the occurrence of anaphora in FU Qs, we have indications that the previous system answer ( $A_1$ ) plays an important role; the number of detected anaphora in the FU Q increased by 61% (relative) when previous system answers were considered as the possible location for antecedents. We see this contextual dependency of FU Qs as an indication that users do take the answers to their previous question into account when formulating a FU Q; it is thus essential to consider this fact also in a new IQA evaluation campaign that goes beyond the context questions task, from which the TREC questions described in this paper are taken. Further research should investigate how to treat different kinds of web search engine results as system answers, and explore to what extent FU Qs refer to such previous search results in a way similar to more traditional IQA systems.

We believe that realistic IQA dialogues like the BoB data described in this paper can serve as a basis for studying, modeling and predicting user topic shifting behavior, particularly with methods based on Machine Learning (e.g., (Kirschner et al., 2009)). Such a study is not possible using artificial IQA data, because even in the case of data originating from a Wizard-of-Oz experiment such as Bertomeu, topic shifts will be to a large extent determined by the user’s particular task when conducting the experiment. On the other hand, empirical and supervised Machine Learning-based approaches are facilitated by the easy availability of

rather large collections of realistic IQA dialogue data, e.g., in the form of dialogue snippets, as we have proposed in this paper. We would hope to see further realistic IQA dialogue collection efforts, possibly in other languages or domains, and the coordinated release of all resulting dialogue data sets to the IQA research community.

## 6. References

- Raffaella Bernardi, Manuel Kirschner, and Zorana Ratkovic. 2010. Context fusion: The role of discourse structure and centering theory. In *Proceedings of LREC 2010*, Malta.
- Nuria Bertomeu. 2008. *A Memory and Attention-Based Approach to Fragment Resolution and its Application in a Question Answering System*. Ph.D. thesis, Department of Computational Linguistics, Saarland University.
- Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162, Stanford, California.
- John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrikari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishede. 2000. Issues, tasks and program structures to roadmap research in question & answering (Q&A).
- Joyce Y. Chai and Rong Jin. 2004. Discourse structure for context question answering. In *Proc. of the HLT-NAACL 2004 Workshop on Pragmatics in Question Answering*, Boston, MA.
- Hoa Trang Dang, Diane Kelly, and Jimmy Lin. 2007. Overview of the TREC 2007 question answering track. In *Proc. of the 16th Text REtrieval Conference*.
- Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proc. of LREC 2006*, Genoa, Italy.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Mijail Kabadjov. 2007. *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification*. Ph.D. thesis, University of Essex.
- Diane Kelly and Jimmy Lin. 2007. Overview of the trec 2006 ciqa task. *SIGIR Forum*, 41(1):107–116.
- Diane Kelly, Paul B. Kantor, Emile L. Morse, Jean Scholtz, and Ying Sun. 2006. User-centered evaluation of interactive question answering systems. In *Proc. of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 49–56, New York, NY.
- Manuel Kirschner, Raffaella Bernardi, Marco Baroni, and Le Thanh Dinh. 2009. Analyzing Interactive QA dialogues using Logistic Regression Models. In *Proc. of AI\*IA*, Reggio Emilia, Italy.
- Manuel Kirschner. 2010. *The Structure of Real User-System Dialogues in Interactive Question Answering*. Ph.D. thesis, Free University of Bozen-Bolzano, Italy.
- Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.
- Mark T. Maybury, editor. 2003. *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium, Stanford University, Stanford, CA, USA*. AAAI Press.
- Ruslan Mitkov, Branimir K. Boguraev, John I. Tait, and Martha Palmer, editors. 2009. *Journal of Natural Language Engineering. Special Issue on Interactive Question Answering*, volume 15. Cambridge University Press.
- Massimo Poesio and Mijail Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proc. of the 4th International Conference on Language Resources And Evaluation (LREC)*, Lisbon, Portugal.
- Zorana Ratkovic. 2009. Deep analysis in iqa: evaluation on real users’ dialogues. Master’s thesis, European Masters Program in Language and Communication Technologies.
- Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. 1998. Analysis of a very large altavista query log. Technical Report 14, Compaq Systems Research Centre, Palo Alto, CA.
- Tomek Strzalkowski and Sanda Harabagiu, editors. 2006. *Advances in Open Domain Question Answering*, volume 32 of *Text, Speech and Language Technology*. Springer Netherlands.
- Tomek Strzalkowski. 2006. The future: Interactive, collaborative information systems. Slides presented at HLT-NAACL 2006 Workshop on Interactive Question Answering.
- Mingyu Sun and Joyce Y. Chai. 2007. Discourse processing for context question answering based on linguistic knowledge. *Know.-Based Syst.*, 20(6):511–526.
- Ellen M. Voorhees. 2001. Overview of the TREC 2001 question answering track. In *Proc. of the 10th Text REtrieval Conference*.
- Ellen M. Voorhees. 2004. Overview of the TREC 2004 question answering track. In *Proc. of the 13th Text REtrieval Conference*.
- Fan Yang, Junlan Feng, and Giuseppe Di Fabbrizio. 2006. A data driven approach to relevancy recognition for contextual question answering. In *Proc. of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 33–40, New York City, NY.

	BoB	TREC'01	TREC'04	Bertomeu
Number $Q_2$ s (= nr. snippets)	1,522	32	221	1,052
pers-pro (GuiTAR)	71 (5%)	5 (16%)	48 (22%)	110 (10%)
pn (GuiTAR)	115 (8%)	2 (6%)	4 (2%)	159 (15%)
the-np (GuiTAR)	9 (1%)	0	5 (2%)	22 (2%)
the-pn (GuiTAR)	0	1 (3%)	0	4 (0%)
Anaphora total (GuiTAR)	195 (13%)	8 (25%)	57 (26%)	295 (28%)

Table 7: Comparison of **automatic anaphora counts** across data sets, without considering previous answer ( $A_1$ ). Percentages are out of total  $Q_2$ s.

	Anaphora in $Q_2$ , considering $A_1$	Anaphora in $Q_2$ , <b>without</b> considering $A_1$	Relative change (additional anaphora through $A_1$ )
Personal pronoun (pers-pro)	130 (9% of 1,522)	71 (5%)	+ 59 (+ 83%)
Proper noun (pn)	140 (9%)	115 (8%)	+ 25 (+ 22%)
Definite NP (the-np)	43 (3%)	9 (1%)	+ 34 (+ 378%)
Anaphora total	313 (21%)	195 (13%)	+ 118 (+ 61%)

Table 8: Automatic anaphora counts (using GuiTAR) in 1,522 BoB  $Q_2$ s, with/without **considering previous answer** ( $A_1$ )

	TREC'01	TREC'04
Number $Q_2$ s (= nr. snippets)	32	221
Anaphora total (automatic)	8 (25%)	57 (26%)
Anaphora total (manual)	22 (69%)	173 (78%)

Table 9: Comparing automatic anaphora counts (using GuiTAR) with **manual anaphora counts** in TREC data

	BoB	BoB, TC $Q_2$ s only	TREC'01	TREC'04	Bertomeu
$Q_1.A_1$	0.08	0.12	–	–	0.09
$Q_1.Q_2$	0.24	0.39	0.25	0.31	0.30
$Q_1.A_2$	0.08	0.12	–	–	0.05
$A_1.Q_2$	0.08	0.09	–	–	0.06
$A_1.A_2$	0.14	0.16	–	–	0.14
$Q_2.A_2$	0.18	0.17	–	–	0.09

Table 10: Comparison of average inter-utterance **string similarities** across data sets

	BoB	BoB, TC $Q_2$ s only	TREC'01	TREC'04	Bertomeu
Number $Q_2$ s (= nr. snippets)	1,522	250	32	221	1,052
center.Reference( $Q_2 \rightarrow A_2$ )	3%	4%	–	–	3%
center.Forward( $Q_1 \rightarrow A_2$ )	39%	53%	–	–	15%
center.Transition( $Q_1 \rightarrow Q_2$ ): continue	8%	17%	16%	40%	14%
center.Transition( $Q_1 \rightarrow Q_2$ ): retain	2%	3%	0%	1%	2%
center.Transition( $Q_1 \rightarrow Q_2$ ): smoothShift	1%	2%	0%	3%	1%
center.Transition( $Q_1 \rightarrow Q_2$ ): roughShift	89%	78%	84%	55%	82%
ConstraintRefinement ( $Q_1 \rightarrow Q_2$ )	4%	7%	3%	8%	7%
ParticipantShift ( $Q_1 \rightarrow Q_2$ )	4%	8%	3%	6%	9%
TopicExploration ( $Q_1 \rightarrow Q_2$ )	28%	48%	28%	52%	42%

Table 11: Comparison of **Dialogue and Discourse features** across data sets in proportions of all  $Q_2$ s (= snippets). Without considering previous answer  $A_1$ .