Towards an Empirically Motivated Typology of Follow-Up Questions: The Role of Dialogue Context

Manuel Kirschner and **Raffaella Bernardi**
KRDB Centre, Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
{kirschner,bernardi}@inf.unibz.it

Abstract

A central problem in Interactive Question Answering (IQA) is how to answer Follow-Up Questions (FU Qs), possibly by taking advantage of information from the dialogue context. We assume that FU Qs can be classified into specific types which determine if and how the correct answer relates to the preceding dialogue. The main goal of this paper is to propose an empirically motivated typology of FU Qs, which we then apply in a practical IQA setting. We adopt a supervised machine learning framework that ranks answer candidates to FU Qs. Both the answer ranking and the classification of FU Qs is done in this framework, based on a host of measures that include shallow and deep inter-utterance relations, automatically collected dialogue management meta information, and human annotation. We use Principal Component Analysis (PCA) to integrate these measures. As a result, we confirm earlier findings about the benefit of distinguishing between topic shift and topic continuation FU Qs. We then present a typology of FU Qs that is more fine-grained, extracted from the PCA and based on real dialogue data. Since all our measures are automatically computable, our results are relevant for IQA systems dealing with naturally occurring FU Qs.

1 Introduction

When real users engage in written conversations with an Interactive Question Answering (IQA) system, they typically do so in a sort of dialogue rather than asking single shot questions. The questions' context, i.e., the preceding interactions, should be useful for understanding Follow-Up Questions (FU Qs) and helping the system pinpoint the correct answer. In previous work (Kirschner et al., 2009; Bernardi et al., 2010; Kirschner, 2010), we studied how dialogue context should be considered to answer FU Qs. We have used Logistic Regression Models (LRMs), both for learning which aspects of dialogue structure are relevant to answering FU Qs, and for comparing the accuracy with which the resulting IQA systems can correctly answer these questions. Unlike much of the related research in IQA, which used artificial collections of user questions, our work has been based on real user-system dialogues we collected via a chatbot-inspired help-desk IQA system deployed on the web site of our University Library.

Previously, our experiments used a selection of shallow (Kirschner et al., 2009) and deep (Bernardi et al., 2010) features, all of which describe specific relations holding between two utterances (i.e., user questions or system answers). In this paper we present additional features derived from automatically collected dialogue meta-data from our chatbot's dialogue management component. We use Principal Component Analysis (PCA) to combine the benefits of all these information sources, as opposed to using only certain hand-selected features as in our previous work.

The main goal of this paper is to learn from data a new typology of FU Qs; we then compare it to an existing typology based on hand-annotated FU Q types, as proposed in the literature. We show how this new typology is effective for finding the correct answer to a FU Q. We produce this typology by analyzing the main components of the PCA.

This paper presents two main results. A new, empirically motivated typology of FU Qs confirms earlier results about the practical benefit of distinguishing between topic continuation and topic shift FU Qs, which are typically based on hand annotation. We then show that we can do without such hand annotations, in that our fully automatic,

on-line measures – which include automatically collected dialogue meta-data from our chatbot's dialogue manager – lead to better performance in identifying correct answers to FU Qs.

In the remainder of this paper, we first review relevant previous work concerning FU Q typologies in IQA. Section 3 then introduces our collection of realistic IQA dialogues which we will use in all our experiments; the section includes descriptions of meta information in the form of dialogue management features and post-hoc human annotations. In Section 4 we introduce our experimental framework, based on inter-utterance features and LRMs. Our experimental results are presented in Section 5, which is followed by our conclusions.

2 Related work

Much of previous work on dialogue processing in the domain of contextual or interactive Question Answering (QA) (Bertomeu, 2008; van Schooten et al., 2009; Chai and Jin, 2004; Yang et al., 2006) has been based on (semi-)artificially devised sets of context questions. However, the importance of evaluating IQA against *real* user questions and the need to consider preceding system answers has already been emphasized (Bernardi and Kirschner, 2010). The corpus of dialogues we deal with consists of real logs in which actual library users were conversing (by typing) with a chat-bot to obtain information in a help-desk scenario.

(Yang et al., 2006) showed that shallow similarity features between a FU Q and the preceding utterances are useful to determine whether the FU Q is a continuation of the on-going topic ("topic continuation"), or it is a "topic shift". The authors showed that recognizing these two basic types of FU Qs is important for deciding which context fusion strategies to employ for retrieving the answer to the FU Q. (Kirschner et al., 2009) showed how shallow measures of lexical similarity between questions and answers in IQA dialogues are as effective as manual annotations for distinguishing between these basic FU Q types. However, that earlier work was based on a much smaller set of dialogue data than we use in this paper, making for statistically weaker results. (Bernardi et al., 2010) improved on this approach by increasing the data set, and adding "deep" features that quantify text coherence based on different theories of dialogue and discourse structure. However, FU Q classification was performed using either single, hand-selected shallow or deep features, or a hand-selected combination of one shallow and one deep feature. In this paper, we adopt the most promising measures of similarity and coherence from the two aforementioned papers, add new features based on automatically collected dialogue management meta-data, and combine all this information via Principal Component Analysis (PCA). By using PCA, we circumvent the theoretical problem that potentially multicollinear features pose to our statistical models, and at the same time we have a convenient means for inducing a new typology of FU Qs from our data, by analyzing the composition of the principal components of the PCA.

More fine-grained typologies of FU Qs have been suggested, and different processing strategies have been proposed for the identified types. In this paper, we start from our own manual annotation of FU Qs into four basic classes, as suggested by the aforementioned literature (Bertomeu, 2008; van Schooten et al., 2009; Sun and Chai, 2007). We then compare it to our new PCA-based FU Q typology.

3 Data

We now introduce the set of IQA dialogue data which we will use in our experiments. For the purpose of calculating inter-utterance features within these user-system interactions – as described in Section 4.4 – we propose to represent utterances in terms of *dialogue snippets*. A dialogue snippet, or *snippet* for short, contains a FU Q, along with a 2-utterance window of the preceding dialogue context. In this paper we use a supervised machine learning approach for evaluating the correctness of a particular answer to a FU Q; we thus represent also the answer candidate as part of the snippet. Introducing the naming convention we use throughout this paper, a snippet consists of the following four successive utterances: Q_1, A_1, Q_2, and A_2. The FU Q is thus referred to as Q_2.

The data consists of 1,522 snippets of 4-turn human-machine interactions in English: users ask questions and the system answers them. The data set was collected via the Bolzano Bot (BoB) web application that has been working as an on-line virtual help desk for the users of our University Library since October 2008.[1] The snippets were

---

[1] `www.unibz.it/library`. More information on the BoB dialogue corpus: `bob.iqa-dialogues.net`.

extracted from 916 users' interactions.

Table 3 shows three example dialogue snippets with correct A_1 and A_2; these examples are meant to give an idea of the general shape of the BoB dialogue data. In the third example snippet, A_1 and A_2 actually contain clickable hyperlinks that open an external web-site. We represent them here as dots in parentheses.

Our library domain experts manually checked that each FU Q was either correctly answered in the first place by BoB, or they corrected BoB's answer by hand, by assigning to it the correct answer from BoB's answer repository. In this way, the dialogue data contain 1,522 FU Qs, along with their respective contexts (Q_1 and A_1) and their *correct* answers (A_2). The resulting set of correct A_2s contains 306 unique answers.[2]

The BoB dialogue data also contain two levels of meta information that we will use in this paper. On the one hand, we have automatically collected dialogue meta-data from BoB's dialogue manager that describe the internal state of the BoB system when a FU Q was asked; this information is described in Section 4.2. On the other hand, 417 of the 1,522 FU Qs were hand-annotated regarding FU Q type, as described in Section 4.3.

4 Model

Our goal is, given a FU Q (Q_2 in our dialogue snippets), to pick the best answer from the fixed candidate set of 306 A_2s, by assigning a score to each candidate, and ranking them by this score. Different FU Q types might require different answer picking strategies. Thus, we specify both A_2 *(identification) features*, aiming at selecting the correct A_2 among candidates, and *context (identification) features*, that aim at characterizing the context. The A_2 identification features measure the similarity or coherence between an utterance in the context (e.g., Q_2) and a candidate A_2. Context features measure the similarity or coherence between pairs of utterances in the context (e.g., Q_1 and Q_2). They do not provide direct information about A_2, but might cue a special context (say, an instance of topic shift) where we should pay more attention to different A_2 identification features (say, less attention to the relation between

Q_2 and A_2, and more to the one between A_1 and A_2).

We implement these ideas by estimating a generalized linear model from training data to predict the probability that a certain A_2 is correct given the context. In this model, we enter A_2 features as main effects, and context features in interactions with the former, allowing for differential weight assignment to the same A_2 features depending on the values of the context features.

4.1 Logistic Regression

Logistic regression models (LRMs) are generalized linear models that describe the relationship between features (independent variables) and a binary outcome (Agresti, 2002). LRMs are closely related to Maximum Entropy models, which have performed well in many NLP tasks. A major advantage of using logistic regression as a supervised machine learning framework (as opposed to other, possibly better performing approaches) is that the learned coefficients are easy to interpret and assess in terms of their statistical significance. The logistic regression equations specify the probability for a particular answer candidate A_2 being correct, depending on the β coefficients (representing the contribution of each feature to the total answer correctness score), and the feature values $x_1, \ldots, x_k$. In our setting, we are only interested in the *rank* of each A_2 among all answer candidates, which can be easily and efficiently calculated through the linear part of the LRM: score $= \beta_1 x_1 + \ldots + \beta_k x_k$.

FU Q typology is implicitly modeled by *interaction* terms, given by the product of an A_2 feature and a context feature. An interaction term provides an extra β to assign a differential weight to an A_2 feature depending on the value(s) of a context feature. In the simplest case of interaction with a binary 0-1 feature, the interaction β weight is only added when the binary feature has the 1-value.

As described in (Kirschner, 2010), we estimate the model parameters (the beta coefficients $\beta_1, \ldots, \beta_k$) using maximum likelihood estimation. Moreover, we put each model we construct under trial by using an iterative backward elimination procedure that keeps removing the least significant predictor from the model until a specific stopping criterion that takes into account the statistical goodness of fit is satisfied. All the results

---

[2]Many of the 306 answer candidates overlap semantically. This is problematic, given that our evaluation approach assumes exactly *one* candidate to be correct, while all other 305 answers to be wrong. In this paper, we shall accept this fact, for the merit of simplicity.

we report below are obtained with models that underwent this trimming procedure.

There is a potential pitfall when using multiple regression models such as LRMs with multi-collinear predictors, i.e., predictors that are inter-correlated, such as our alternative implementations of inter-utterance string similarity. In such situations, the model may not give valid results about the importance of the individual predictors. In this paper, we use PCA to circumvent the problem by combining potentially multicollinear predictors to completely uncorrelated PC-based predictors.

In the following three sections, we describe the different types of information that are the basis for our features.

4.2 BoB dialogue management meta-data

When BoB interacts with a user, it keeps log files of the IQA dialogue. First of all, these logs include a timed protocol of user input and BoB's responses: the user and system utterances are the literal part of the information. On the other hand, BoB also logs two dimensions of meta information, both of which are based on BoB's internal status of its *dialogue management* routine. This routine is based on a main *initiative-response* loop, mapping user input to some canned-text answer, where the user input should be matched by (at least) one of a set of hand-devised regular expression question patterns.

Sub-dialogues Whenever BoB asks a system-initiated question, the main loop is suspended, and the system goes into a *sub-dialogue* state, where it waits for a specific response from the user – typically a short answer indicating the user's choice about one of the options suggested by BoB. The next user input is then matched against a small number of regular expression patterns specifically designed for the particular system-intiative question at hand. Depending on this user input, the sub-dialogue can:

`Continue:` the user input matched one of the regular expression patterns intended to capture possible user choices

`Break:` the user broke the sub-dialogue by entering something unforeseen, e.g., a new question

The first two parts of Table 4 give an overview of the statistics of BoB's dialogue management-based meta information concerned with sub-dialogue status. Besides *continue* and *break*, for Q_1 we consider also a third, very common case that a user question was not uttered in a sub-dialogue setting at all. Note that we excluded from our data collection all those cases where Q_2 continues a sub-dialogue from our collection of IQA dialogues, since we do not consider such Q_2s as FU Qs, as they are highly constrained by the previous dialogue.

Apology responses The third part of Table 4 gives statistics of whether a particular system response A_1 was an apology message stating that BoB did not understand the user's input, i.e., none of BoB's question patterns matched the user question.

4.3 Manual dialogue annotation

We now turn to the meta information in BoB dialogue data that stems from post-hoc human annotation. For a portion of BoB's log files, we added up to two additional levels of meta information, by annotating the log files after they were collected.[3]

The following paragraphs explain the individual levels of annotation by giving the corresponding annotator instructions; Table 5 contains an overview of the corresponding features. First of all, we annotated FU Qs with their FU Q type. Our choice of the particular four levels of the `FUQtype` feature was influenced by the following literature literature: from (De Boni and Manandhar, 2005) and (Yang et al., 2006) we adopted the distinction between topic shift and topic continuation, while from (Bertomeu et al., 2006) we took the notions of rephrases and context dependency. Our annotation scheme is described in Figure 1; note that topic continuations have three sub-types, which are spelled out below.

`FUQtype = isTopicShift:` marks a FU Q as a *topic shift* based on an intuitive notion of whether the FU Q "switches to something completely different".

`FUQtype = isRephrase:` marks whether the FU Q is an attempt to re-formulate the same question. The FU Q could be a literal repetition of the previous question, or it could be a rephrasing.

`FUQtype = isContextDepentFUQ:` marks whether the FU Q needs to be considered along with some information provided by

---

[3]All annotations were performed by either one of the authors.

the dialogue context in order to be correctly understood.

FUQtype = isFullySpecifiedFUQ:
marks whether the FU Q does not need any information from the dialogue context in order to be correctly understood.

The second level of hand-annotation concerns a manual check of the correctness of A_1. It is available for 1,179 of our 1,522 snippets.

A1.isAnswer.correct: marks whether the system response is correct for the given question.

A1.isApology.correct: marks whether BoB's apology message is correct for the given question.

4.4 Shallow/deep inter-utterance relations

We exploit shallow features, which measure the similarity between two utterances within a snippet, and deep features, which encode coherence between two utterances based on linguistic theory. For each feature we will use names encoding the utterances involved; e.g., `distsim.A1.Q2` stands for the Distributional Similarity feature calculated between A_1 and Q_2.

Shallow features The detailed description of all the shallow features we used in our experiments can be found in (Kirschner et al., 2009). The intuition is that a high similarity between Q and A tends to indicate a correct answer, while in the case of high similarity between the dialogue context and the FU Q, it indicates a "topic continuation" FU Q (as opposed to a "topic shift" FU Q), and thus helps discriminating these two classes of FU Qs.

`Lexical Similarity (lexsim):` If two utterances share some terms, they are similar; the more *discriminative* the terms they share, the more similar the utterances. Implements a TF-IDF-based similarity metric. `Distributional Similarity (distsim.svd):` Two utterances are similar not only if they share the same terms, but also if they share similar terms (e.g., *book* and *journal*). Term similarity is estimated on a corpus, by representing each content word (noun, verb, adjective) as a vector that records its corpus co-occurrence with other content words within a 5-word span. `Action sequence (action):` Based on the notion that in our helpdesk setting we are dealing with task-based dia-

logues, which revolve around library-related actions (e.g., "borrow", "search"). The action feature indicates whether two utterances contain the same action.

Deep features These features encode different theories of discourse and dialogue coherence. Refer to (Bernardi et al., 2010) for a full description of all deep features we used experimentally, along with more details on the underlying linguistic theories, and our implementation choices for these features.

We introduce a four-level feature, `center`, that encodes the four transitions holding between adjacent utterances that Centering Theory describes (Brennan et al., 1987; Grosz et al., 1995). Somewhat differently from that classic theory, (Sun and Chai, 2007) define the transitions depending on whether both the head and the modifier of the Noun Phrases (NP) representing the *preferred centers*[4] are continued (`cont`) or switched (rough shift: `roughSh`) between Q_1 and Q_2. The remaining two transitions are defined in similar terms.

4.5 PCA-based context classification features

Principal Component Analysis (PCA) (Manly, 2004) is a statistical technique for finding patterns in high-dimensional data, or for reducing their dimensionality. Intuitively, PCA rotates the axes of the original data dimensions in such a way that few of the new axes already cover a large portion of the variation in the data. These few new axes are represented by the so-called principal components (PCs). We employ this technique as a tool for combining a multitude of potentially multi-collinear predictors for context classification, i.e., all predictors that involve Q_2 and some preceding utterance. In our experiments we will also want to look at the correlations of each of the top PCs with the original context classification features; these correlations are called *loadings* in PCA. We experiment with the following three versions of PCA:

PCA$_A$: without BoB dialogue management meta-data features PCA performed over all context classification features of the shallow and deep types described in Section 4.4.

---

[4]Centers are noun phrases. The syntactic structure of a noun phrase comprises a *head noun*, and possibly a *modifier*, e.g., an adjective. We use a related approach, described in (Ratkovic, 2009), to identify the *preferred center* of each question.

PCA$_B$: with BoB dialogue management meta-data features PCA$_A$ plus BoB's dialogue-management meta-data features (Section 4.2).

PCA$_C$: with BoB dialogue management meta-data features and manual A_1 correctness check PCA$_B$ plus additional manual annotation of A_1 correctness (Section 4.3).

5 Evaluation

We employ a standard 10-fold cross-validation scheme for splitting training and prediction data. We assess our LRMs by comparing the ranks that the models assign to the gold-standard correct A_2 candidate (i.e., the single A_2 that our library domain experts had marked as correct for each of the 1,522 FU Qs). To determine whether differences in A_2 ranking performance are significant, we consult both the paired t-test and the Wilcoxon signed rank test about the difference of the 1,522 ranks.

5.1 Approximating hand-annotated FU Q types with PCA-based features

We begin the evaluation of our approach by exploring the value of the hand-annotation-based FU Q type as cues for expressing the relevance and topical relatedness of that particular FU Q's dialogue context.

For this purpose, we use the subset of 417 dialogue snippets which we annotated with the `FUQtype` feature described in the first half of Table 5. Figure 1 depicts our FU Q type taxonomy, and the distribution of the four types in our data.

First of all, for this hand-annotated subset of dialogue snippets, we try to improve the A_2 ranking results of a "main effects only" **baseline** LRM, i.e., a model which does not distinguish between different FU Q types. This baseline model was proposed in earlier work (Kirschner et al., 2009). We tried the following features as interaction term(s) in our models, one after the other: whether the hand-annotated `FUQType` feature indicates a topic shift or not; the full four levels of `FUQType`; a linear combination of the top five PCs of each of the three PCA feature sets introduced in Section 4.5. After applying our automatic predictor elimination routine described in Section 4.1 and evaluating the A_2 ranking results of each of these models, none of the interactive models significantly outperform our baseline. PCA-based context classification using only fully automatic BoB meta information features (PCA$_B$ in Section 4.5) results

in the largest improvement over baseline; however, this improvement does not reach statistical significance, most likely due to the small data set of only 417 cases. Still, using the hand-annotated FU Q type feature `FUQType`, we can visualize how the top PCs cluster the 417 FU Qs, and how this clustering mirrors some of the distinctions of manually assigned FU Q types: see Figure 2. E.g., plotting the FU Qs along their PC1 and PC2 values seems to mimic the annotator's distinction between topic shift FU Qs and the other three FU Q types. The other pairs of PCs also appear to show certain clusters. Overall, the automatic context classification features that served as input to the PCA are useful for describing different context-related behaviors of different FU Qs.

5.2 Optimizing A_2 ranking scores using PCA-based features

Having shown the usefulness (in terms of assigning high ranks to the gold-standard correct A_2) of FU Q classification via a PCA-based combination of purely automatic context classification features, we can now consider the full sample of 1,522 dialogue snippets described in Section 3, for which we do not in general possess manual FU Q type annotations.

The first row of Table 1 shows the A_2 ranking results of our baseline LRM. In the remainder of the table, we compare this baseline model to three different models which use a linear combination of different versions of the top five PCs as interaction terms. The three versions (A, B and C) were introduced in Section 4.5.

5.3 Analysis of PC-based context features

The main goal of this paper is to devise an empirically motivated typology of FU Qs, under consideration of automatically collected dialogue management meta information. We then want to show how this new typology is effective for finding the correct answer to a specific FU Q, in that for the given FU Q it indicates the relevance and topical relatedness of the question's particular dialogue context. In Section 5.2 we have seen how all PCA-based context classification features perform clearly better than a non-interactive baseline model; more specifically, the top five PCs from the PCA$_B$ scheme yield significantly better A_2 ranking results than the PCA$_A$ scheme which does not consider BoB dialogue management meta-data features. Based on these results, we now look in

Model ID	Interaction terms	Mean rank correct A_2	Median rank correct A_2	Standard dev.	p (Paired t-test)	p (Wilcoxon signed rank)
baseline	none	48.72	14	69.35		
PCA_A	$PC1 + \ldots + PC5$	44.25	12	64.58	< 0.0001	< 0.0001
PCA_B	$PC1 + \ldots + PC5$	**42.72**	**12**	62.53	0.0006	0.0087
PCA_C	$PC1 + \ldots + PC5$	42.87	12	62.94	not sig.	not sig.

Table 1: Improving ranking of correct A_2 (out of 306 answer candidates) with different PCA-based interaction terms. Significance tests of rank differences wrt. result in preceding row.

more detail at the relevance of the top five PC features in PCA_B, and at their most important *loadings*, i.e., the original context classification features that are most highly correlated with the value of each particular PC. After running our predictor elimination routine, the corresponding LRM has kept three of these five top PCs as interaction terms: PC1, PC2 and PC5. Table 2 describes the top three positive and top three negative loadings of these PCs. The table also shows how in model PCA_B, each of the interaction terms corresponding to the three PCs influences the score that is calculated for every A_2 candidate, either positively or negatively.

Interpreting the results of Table 2 on a high, dialogue-specific level, we draw the following conclusions:

PC1 seems to capture a rather general distinction of topic shift versus topic continuation. A FU Q with high lexical similarity to the preceding utterances (i.e., a "topic continuation") should preferably get an A_2 with higher lexical similarity with respect to both A_1 and Q_2. In this context, "topic shift" is partly described by a feature from Centering Theory, and two of BoB's dialogue management meta-data features.

PC2 shows relatively weak *positive* correlations with any context classification features. On the negative end, PC2 seems to describe a class of FU Qs that are uttered after a Q_1 that did neither continue nor exit a sub-dialogue. Also, A_1 was a regular system answer (as opposed to an apology message by BoB). Such FU Qs can thus be interpreted as "single shot" questions that a user poses after their previous question was already dealt with in A_1. Because of the negative loadings, the value of PC2 becomes negative, resulting in the *avoidance* of any A_2 that is highly similar to the preceding A_1.

PC5 distinguishes FU Qs that are mostly related to the previous answer from those that are more related to the previous question. Depending on whether PC5 turns positive or negative, A_2s are preferred that are more similar to A_1 or Q_2, respectively. $Q_1.Q_2$ similarity is determined by both lexical similarity and Centering Theory features.

6 Conclusion

In this paper we have experimentally explored the problem of FU Q types and their corresponding answer identification strategies. The first result is that our hand-annotated FU Q types did not significantly improve Q_2 answering performance (for the annotated sub-set of 417 snippets). We attribute this negative result in part to the difficulty of the 4-level FU Q type annotation task. On the other hand, we believe it is encouraging that with purely automatic features for context classification, combined through PCA, we significantly outperformed our baseline. Adding BoB's dialogue management meta information – which is also automatically available when using our dialogue collection scheme – for context classification helped improve the scores even further. We analyzed the top loadings of three PCs that our best-performing LRM uses for FU Q type classification. We used PCA both for circumventing the problem of multicollinear predictors in LRM, and as a diagnostic tool to analyze the most important components of automatically combined FU Q classification features. Finally, a potentially difficult and cumbersome manual annotation of the correctness of the previous system answer A_1 did not improve A_2 ranking performance.

References

Alan Agresti. 2002. *Categorical Data Analysis.* Wiley-Interscience, New York.

Raffaella Bernardi and Manuel Kirschner. 2010. From

PC1		PC2		PC5	
0.33	`distsim.Q1.Q2`	0.05	`Q1.bob.contSubdial`	0.45	`distsim.A1.Q2`
0.26	`distsim.A1.Q2`	0.04	`Q2.center.roughSh`	0.31	`A1.bob.isApology`
0.26	`action.Q1.Q2`	0.02	`Q2.bob.breakSubdial`	0.29	`lexsim.A1.Q2`
$\vdots$		$\vdots$		$\vdots$	
-0.13	`A1.bob.isApology`	-0.22	`A1.bob.isAnswer`	-0.18	`lexsim.Q1.Q2`
-0.15	`Q2.bob.noSubdial`	-0.30	`Q2.bob.noSubdial`	-0.23	`Q2.center.cont`
-0.22	`Q2.center.roughSh`	-0.31	`Q1.bob.noSubdial`	-0.26	`A1.bob.isAnswer`

INFLUENCE ON A_2 SELECTION IN MODEL PCA$_B$

pos for each A_2 similar to Q_2	**pos** for each A_2 similar to A_1	**pos** for each A_2 similar to A_1
pos for each A_2 similar to A_1		**neg** for each A_2 similar to $Q2$

Table 2: Strongest loadings for the three PCs retained as interaction terms in Model PCA$_B$, and indication of each PC's positive/negative influence on lexical similarity-based A_2 selection features

artificial questions to real user interaction logs: Real challenges for interactive question answering systems. In *Proc. of Workshop on Web Logs and Question Answering (WLQA'10)*, Valletta, Malta.

Raffaella Bernardi, Manuel Kirschner, and Zorana Ratkovic. 2010. Context fusion: The role of discourse structure and centering theory. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Núria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 2006. Contextual phenomena and thematic relations in database QA dialogues. In *Proc. of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 1–8, New York, NY.

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.

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.

Marco De Boni and Suresh Manandhar. 2005. Implementing clarification dialogues in open domain question answering. *Journal of Natural Language Engineering*, 11(4):343–361.

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.

Manuel Kirschner, Raffaella Bernardi, Marco Baroni, and Le Thanh Dinh. 2009. Analyzing Interactive QA dialogues using Logistic Regression Models. In *Proc. of XIth International Conference of the Italian Association for Artificial Intelligence (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.

Bryan F. J. Manly. 2004. *Multivariate Statistical Methods: A Primer*. Chapman & Hall/CRC, 3rd edition.

Zorana Ratkovic. 2009. Deep analysis in iqa: evaluation on real users' dialogues. Master's thesis, European Masters Program in Language and Communication Technologies.

Mingyu Sun and Joyce Y. Chai. 2007. Discourse processing for context question answering based on linguistic knowledge. *Know.-Based Syst.*, 20(6):511–526.

Boris van Schooten, R. op den Akker, R. Rosset, O. Galibert, A. Max, and G. Illouz. 2009. Follow-up question handling in the IMIX and Ritel systems: A comparative study. *Journal of Natural Language Engineering*, 15(1):97–118.

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.

Q_1	where in the library can i find books
A_1	Most of our holdings are accessible in the open-shelves area. In the repository we only keep some older volumes of journals or the outsize stuff.
Q_2	shelf mark
A_2	Shelf marks are a combination of letters and numbers, which exactly define the location of a book in a Library. If you find a book or a journal in the catalogue, you also see the shelf mark. The shelf marks of the University Library consist of an alphanumeric code (e.g. ES 560 K19). Would you like to know how you search them?
Q_1	When can I return a book?
A_1	Books and other media can be consigned at the information desk during the regular opening hours of the Library. You can also leave them at the book drop located in front of the main entrance of the university building - 24 hours a day.
Q_2	I want to read the newspaper on the computer
A_2	If you haven't got a Library Card you can visit the Library to read books, journals or newspapers. If you want to borrow something or use the other services we offer you should get a Library Card. Do you want to know how it works?
Q_1	i need the opac
A_1	Here (...) is the page with our OPAC for you. You can search the complete holdings of the Library.
Q_2	where is the advisory service
A_2	The University's Advisory Service provides an information service for prospective students who would like to know more about which degree course to study. It also provides a support service for enrolled students during their entire time at the University. You can get further information from this (...) site.

Table 3: Example dialogue snippets with correctly identified A_2

Feature name	Freq.	Description
Q1.bob.contSubdial	7.6%	Q_1 continues system-initiated sub-dialogue
Q1.bob.breakSubdial	9.6%	Q_1 breaks out of system-initiated sub-dialogue
Q1.bob.noSubdial	82.9%	BoB not in sub-dialogue mode when Q_1 was uttered
Q2.bob.breakSubdial	13.6%	Q_2 breaks out of system-initiated sub-dialogue
Q2.bob.noSubdial	86.4%	BoB not in sub-dialogue mode when Q_2 was uttered
A1.bob.isAnswer	75.6%	A_1 is regular answer retrieved by BoB
A1.bob.isApology	24.4%	A_1 is apology message: BoB did not understand

Table 4: BoB dialogue management meta information. Proportions out of those 1,441 of total 1,522 snippets for which this information was logged.

Feature name	Freq.	Description
FUQtype=isTopicShift	40.0% (of 417)	Q_2 is topic shift
FUQtype=isRephrase	19.2% (of 417)	Q_2 is rephrasing of Q_1
FUQtype=isContextDepentFUQ	6.5% (of 417)	Q_2 is context dependent
FUQtype=isFullySpecifiedFUQ	34.3% (of 417)	Q_2 is not context dependent
A1.isAnswer.correct	66.5% (of 1,179)	BoB's regular answer A_1 is correct
A1.isAnswer.false	19.0% (of 1,179)	BoB's regular answer A_1 is false
A1.isApology.correct	1.3% (of 1,179)	BoB's apology message A_1 is correct
A1.isApology.false	13.2% (of 1,179)	BoB's apology message A_1 is false

Table 5: Manual annotation meta information. Proportions out of those sub-sets of total 1,522 snippets with available annotation.
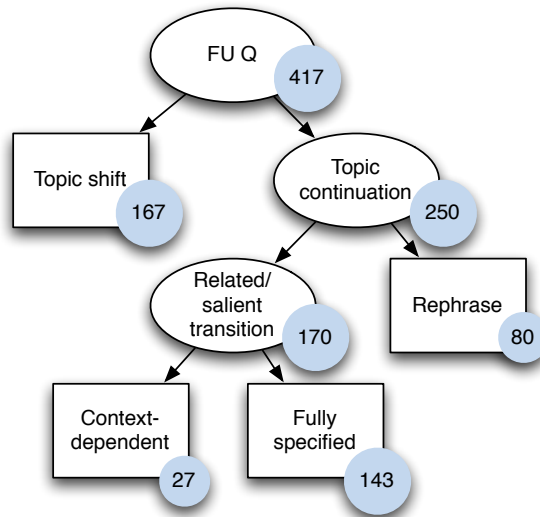
Figure 1: Manual FU Q type annotation scheme, with counts of FU Q types
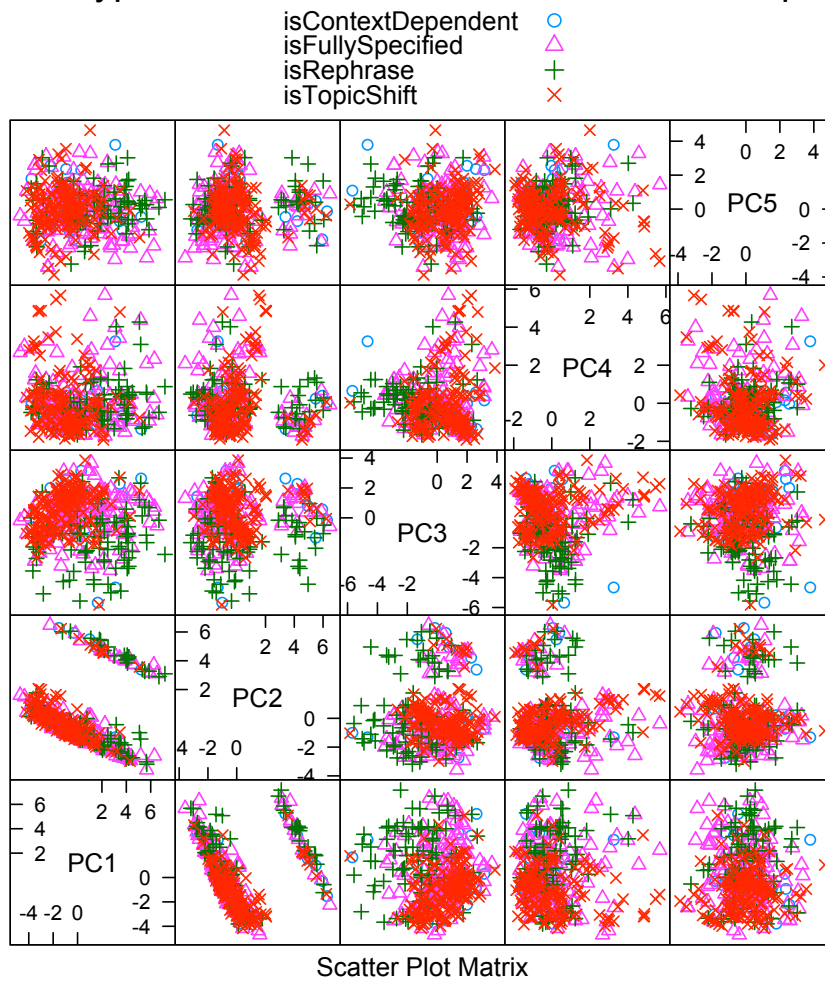

Figure 2: Distribution of hand-annotated FU Q types in PC-based feature space (PCA$_B$)