Exploring Topic Continuation Follow-up Questions using Machine Learning

Manuel Kirschner
KRDB Center
Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
kirschner@inf.unibz.it

Raffaella Bernardi
KRDB Center
Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
bernardi@inf.unibz.it

Abstract

Some of the Follow-Up Questions (FU Q) that an Interactive Question Answering (IQA) system receives are not topic shifts, but rather continuations of the previous topic. In this paper, we propose an empirical framework to explore such questions, with two related goals in mind: (1) modeling the different relations that hold between the FU Q's answer and either the FU Q or the preceding dialogue, and (2) showing how this model can be used to identify the correct answer among several answer candidates. For both cases, we use Logistic Regression Models that we learn from real IQA data collected through a live system. We show that by adding dialogue context features and features based on sequences of domain-specific actions that represent the questions and answers, we obtain important additional predictors for the model, and improve the accuracy with which our system finds correct answers.

1 Introduction

Interactive Question Answering (IQA) can be described as a fusion of the QA paradigm with dialogue system capabilities. While classical QA is concerned with questions posed in isolation, its interactive variant is intended to support the user in finding the correct answer via natural-language dialogue. In an IQA setting, both the system and the user can pose Follow-Up Questions (FU Q). In the second case, whenever an IQA system receives an additional user question (note that this is what we call a *Follow-Up Question* throughout this work), it can either interpret it as being thematically related to a previous dialogue segment (*topic continuation*), or as a shift to some new, unrelated topic (*topic shift*). A definition of thematic relatedness of FU Qs might rely on the elements of the attentional state, i.e., on the objects, properties and relations that are salient before and after processing the user question. Topic continuation FU Qs should be interpreted within the context, whereas topic shift FU Qs have to be treated as first questions and can thus be processed with standard QA technologies. Therefore, a **first task** in IQA is to detect whether a FU Q is a topic shift or a topic continuation (Yang et al., 2006).

To help answering topic continuation FU Qs, an IQA system would need to fuse the FU Q with certain information from the dialogue context (cf. (van Schooten et al., 2009)). Thus, a **second task** in IQA is to understand which turns in the dialogue context are possible locations of such information, and exactly what kind of information should be considered. Knowing that a FU Q concerns the same topic as the previous question or answer, we thus want to study in more detail the way the informational content of questions and answers evolves before/after the FU Q is asked. A model of these so-called *informational transitions* would provide insights into what a user is likely to ask about next in natural coherent human-machine dialogue.

In order to tackle any of the two IQA tasks mentioned above we need IQA dialogues. Most current work on IQA uses the TREC QA data; the TREC QA tracks in 2001 and 2004 included series of context questions, where FU Qs always depended on the context set by an earlier question from the same series. However, these data were constructed artificially and are not representative of actual dialogues from an IQA system (for instance, system answers are not considered at all). Real IQA data yield challenges for an automatic processing approach (Yang et al., 2006). Our work is based on collecting and analyzing IQA dialogues from users of a deployed system.

In this paper, we address the second task introduced above, namely the study of common relations between the answer to a topic continuation FU Q and other turns in the dialogue context. Our collected dialogue data are from the "library help desk" domain. In many of the dialogues, library users request information about a specific library-related action; we are thus dealing with task-oriented dialogues. This work is based on two hypotheses regarding relations holding between the FU Q's answer and the dialogue context. For studying such relations, we want to explore the usefulness of (1) a representation of the library-related action underlying questions and answers, and (2) a representation of the dialogue context of the FU Q.

2 Background

In order to understand what part of the history of the dialogue is important for processing FU Qs, significant results come from Wizard-of-Oz studies, like (Dahlbäck and Jönsson, 1989; Bertomeu et al., 2006; Kirschner and Bernardi, 2007), from which it seems that the immediate linguistic context (i.e., the last user initiative plus the last system response) provides the most information for resolving any context-dependency of the FU Qs. These studies analyzed one particular case of topic continuation FU Q, namely those questions containing reference-related discourse phenomena (ellipsis, definite description or anaphoric pronoun); we assume that the results could be extended to fully specified questions, too.

Insights about the informational transitions within a dialogue come from Natural Language Generation research. (McCoy and Cheng, 1991) provide a list of informational transitions (they call them focus shifts) that we can interpret as transitions based on certain thematic relations. Depending on the conversation's current focus type, they list specific focus shift candidates, i.e., the items that should get focus as a coherent conversation moves along. Since we are interested in methods for interpreting FU Qs automatically, we decided to restrict ourselves to use only the "action" focus type to represent the focus of questions and answers in IQA dialogues. We conjecture that actions form a suitable and robust basis for describing the (informational) meaning of utterances in our class of task-based "help desk" IQA dialogues. Table 1 shows the focus shift candidates for a current focus of type "action". In this work we concentrate on the informational transitions involving *two actions* (i.e., including one of the focus targets listed in bold face in the table).

Node type	Informational transition targets
Action	Actor, object, etc., of the action – any participant (Fillmore) role; purpose (goal) of action, **next action in some sequence, subactions, specializations of the action**

Table 1: Possible informational transition targets for "action" node type (McCoy and Cheng, 1991)

3 Exploring topic continuation FU Qs using Machine Learning

We base our study of topic continuation FU Qs on the two main results described in Section 2: We study snippets of dialogues consisting of four turns, viz. a user question (Q_{-1}), the corresponding system answer (A_{-1}), the FU Q and its system answer (A_0); we use Logistic Regression Models to learn from these snippets (1) which informational (action-action) transitions hold between A_0 and the FU Q or the preceding dialogue, and (2) how to predict whether a specific answer candidate A_0 is correct for a given dialogue snippet.

3.1 Machine learning framework: Logistic Regression

Logistic regression models (Agresti, 2002) are generalized linear models that describe the relationship between features (predictors) and a binary outcome (in our case: answer correctness). We estimate the model parameters (the beta coefficients $\beta_1, \ldots, \beta_k$) that represent the contribution of each feature to the total answer correctness score using maximum likelihood estimation. Note that there is a close relationship to Maximum Entropy models, which have performed well in many 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. The logistic regression equation which predicts the probability for a particular answer candidate A_0 being correct, depending on the learned intercept β_0, the other beta coefficients and the feature values $x_1, \ldots, x_k$ (which themselves depend on a combination of Q_{-1}, A_{-1}, FU Q or A_0) is:

$$\mathrm{Prob}\{\mathrm{answerCorrect}\} = \frac{1}{1+\exp(-X\hat{\beta})}, \text{ where}$$

$$X\hat{\beta} \;=\; \beta_0 + (\beta_1 x_1 + \ldots + \beta_k x_k)$$

3.2 Dialogue data collection

We have been collecting English human-computer dialogues using BoB, an IQA system which is publicly accessible on the Library's web-site of our university[1]. We see the availability of dialogue data from genuinely motivated visitors of the library web-site as an interesting detail of our approach; our data are less constrained and potentially more difficult to interpret than synthesized dialogues (e.g., TREC context track data), but should on the other hand provide insights into the structure of actual IQA dialogues that IQA systems might encounter. We designed BoB as a simple chatbot-inspired application that robustly matches user questions using regular expression-based question patterns, and returns an associated canned-text answer from a repository of 529. The question patterns and answers have been developed by a team of librarians, and cover a wide range of library information topics, e.g., opening time, lending procedures and different library services. In the context of this work, we use BoB merely as a device for collecting real human-computer IQA dialogues.

As a preliminary step towards automatically modeling action-based informational transitions triggered by FU Qs, we annotated each of the 529 answers in our IQA system's repository with the "library action" that we considered to best represent its (informational) meaning. For this, we had devised a (flat) list of 25 library-related actions by analyzing the answer repository (e.g.: access, borrow, change, deliver). We also added synonymous verbs to our action list, like "obtain" for "borrow". If we did not find any action to represent a system answer, we assigned it a special "generic-information" tag, e.g. for answers to questions like "What are the opening times?".

We base our current study on the dialogues collected during the first four months of the IQA system being accessible via the Library's web site. After a first pass of manually filtering out dialogues that consisted only of a single question, or where the question topics were only non-library-related, the collected corpus consists of 948 user questions (first or FU Qs) in 262 dialogue sessions (i.e., from different web sessions). We hand-annotated the user FU Qs in these dialogues as either "topic continuation" (248 questions), or "topic shift" (150 questions).

The remaining FU Qs are user replies to system-initiative clarification questions, which we do not consider here. For each user question, we marked whether the answer given by the IQA system was correct; in the case of wrong answers, we asked our library domain experts to provide the correct answer that BoB should have returned. However, we only corrected the system answer in those cases where the user did not ask a further FU Q afterwards, as we must not change on-going dialogues.

To get the actual training/test data, we had to further constrain the set of 248 topic continuation FU Qs. We removed all FU Qs that immediately follow a system answer that we considered incorrect; this is because any further FU Q is then uttered in a situation where the user is trying to react to the problematic answer, which clearly influences the topic of the FU Q. Of the then remaining 76 FU Qs, we keep the following representation of the dialogue context: the previous user question Q_{-1} and the previous system answer A_{-1}. We also keep the FU Q itself, and its corresponding correct answer A_0.

Finally, we automatically annotated each question with one or more action tags. This was done by simply searching the stemmed question string for any verb stem from our list of 25 actions (or one of their synonyms); if no action stem is found, we assigned the "generic-information" tag to the question. Note that this simple action detection algorithm for questions fails in case of context-dependent questions where the verb is elided or if the question contains still unknown action synonyms.

[1] `www.unibz.it/library`

3.3 Features

In the machine learning framework introduced above, the model is intended to predict the correctness of a given system answer candidate, harnessing information from the local dialogue context: Q_{-1}, A_{-1}, FU Q and the particular answer candidate A_0. We now introduce different features that relate A_0 to either the FU Q or some other preceding turn of the dialogue. The features describe specific aspects of how the answer candidate relates to the current dialogue. Note that we do not list features relating Q_{-1} and A_0, since our experiments showed no evidence for including them in our models.

tfIdfSimilarityQA, tfIdfSimilarityAA: TF/IDF-based proximity scores (ranging from 0 to 1) between two strings, namely FU Q and A_0, or A_{-1} and A_0, respectively. Based on vector similarity (using the cosine measure of angular similarity) over dampened and discriminatively weighted term frequencies. Definition of the TF/IDF distance: two strings are more similar if they contain many of the same tokens with the same relative number of occurrences of each. Tokens are weighted more heavily if they occur in few documents[2], hence we used a subset of the UK English version of the Web-as-Corpus data[3] to train the IDF scores.

Features based on action sequences. To describe the action-related informational transitions we observe between the FU Q and A_0 and between A_{-1} and A_0, we use two sets of features, both of which are based on *hand-annotated* actions for answers and *automatically assigned* actions for questions. **actionContinuityQA, actionContinuityAA:** simple binary features indicating whether *the same* library action (or one of its synonyms) was identified between the FU Q and A_0, or A_{-1} and A_0, respectively. **lmProbQA, lmProbAA:** encode Statistical Language Model probabilities for action tag sequences, i.e., the probability for A_0 having a certain action, given the action associated with FU Q, or the action of A_{-1}, respectively. The underlying Statistical Language Models are probability distributions over action-action sequences that reflect how likely certain action sequences occur in our IQA dialogues, thus capturing properties of salient action sequences. More technically, we use Witten-Bell smoothed 2-gram statistical language models, which we trained on our action-tagged FU Q data.

4 Results

For the evaluation of the logistic regression model, we proceed as follows. Applying a cross-validation scheme, we split our 76 FU Q training examples randomly into five non-intersecting partitions of 15 (or 16) FU Q (with corresponding Q_{-1}, A_{-1}, and correct A_0) each. To train the logistic regression model, we need training data consisting of a vector of independent variables (the various feature values), along with the binary dependent variable, i.e., "answer correct" or "answer false". We generate these training data by "multiplying out" each training partition's 61 FU Qs (76 minus the held-out test set of 15) with all 529 answer candidates; for each FU Q dialogue snippet used for training, this results in one positive training example (where A_0 is the 1 correct out 529 answer candidates), and 528 negative training examples (for all other answer candidates).

For each of the five training/test partitions, we train a different model. We then evaluate each of these models on their corresponding held-out test set. Following the cross-validation idea through, we also train separate Statistical Language Models on sequences of action tags for each of the five training splits; this ensures that the language model probabilities were never trained on test data. We perform the evaluation in terms of the mean rank that the correct answer A_0 is assigned after ranking all 529 answer candidates (by evaluating the logistic regression equation to yield answer scores).

In the following, we give details of different logistic regression models we experimented with. Initially, we chose a subset from the list of features introduced above. Our goal was to retain as few features as needed to explore our two hypotheses, i.e., whether we can make use of (1) a representation of the FU Q's underlying library action, and/or (2) a representation of the immediate dialogue context. By dropping uninformative features, the result-

[2] Cf. Alias-i's LingPipe documentation `http://alias-i.com/lingpipe/demos/tutorial/stringCompare/read-me.html`

[3] `http://wacky.sslmit.unibo.it`

ing models become simpler and easier to interpret. With this goal in mind, we applied a fast backwards elimination routine that drops uninformative predictors (cf. (Baayen, 2008, p.204)) on the five training data splits. In all five splits, both TF/IDF features turned out to be important predictors; in four of the splits, also lmProbQA was retained. lmProbAA was dropped as superfluous in all but two splits, and actionSimilarityAA was retained only in one. With these results, the set of features we retain for our modeling experiments is: tfIdfSimilarityQA, tfIdfSimilarityAA and lmProbQA.

"Complete" model: tfIdfSimilarityQA, tfIdfSimilarityAA and lmProbQA We estimated logistic regression models on the five cross evaluation training sets using all three features as predictors. Table 2 shows the mean ranks of the correct answer for the five evaluation runs, and an overall mean rank with the average across the five splits.

To illustrate the contribution of each of the three predictors towards the score of an answer candidate, we provide the (relevant linear part of) the learned logistic regression equation for the "complete" model (trained on split 1 of the data). Note that the "answer ranker" evaluates this equation to get a score for an answer candidate A_0.

$$\begin{aligned} X\hat{\beta} &= -8.4 + (9.5 * \mathrm{tfIdfSimilarityQA} + \\ &\quad 4.6 * \mathrm{tfIdfSimilarityAA} + \\ &\quad 1.7 * \mathrm{lmProbQA}) \end{aligned}$$

Reduced model 1: No representation of dialogue context Only the features concerning the FU Q and the answer A_0 (tfIdfSimilarityQA, lmProbQA) are used as predictors in building the logistic regression model. The result is a model that treats every FU Q as a stand-alone question. Across the five models, the coefficient for tfIdfSimilarityQA is roughly five times the size of that for lmProbQA.

Reduced model 2: No action sequences We keep only the two TF/IDF features (tfIdfSimilarityQA, tfIdfSimilarityAA). This model thus does not use any features that depend on human annotation, but only fully automatic features. The coefficient learned for tfIdfSimilarityQA is generally twice as large as that for tfIdfSimilarityAA.

Reduced model 3: No dialogue context, no action sequences Considered as a baseline, this model uses a single feature (tfIdfSimilarityQA) to predict answer correctness, favoring those answer candidates that have the highest lexical similarity wrt. the FU Q.

5 Discussion

In order to better understand the relatively high mean ranks of the correct answer candidates across Table 2, we scrutinized the results of the answer ranker (based on all tests on the "complete" model). The distribution of the ranks of correct answers is clearly skewed; in around half of the 76 cases, the correct answer was actually ranked among the top 20 of the 529 answer candidates. However, the mean correct rank deteriorates badly due to the lowest-ranking third of cases. Analyzing these lowest-ranking cases, it appears that they are often instances of two sub-classes of topic continuation FU Qs: (i) the FU Q is context-dependent, i.e., underspecified or exhibiting reference-related discourse phenomena; (ii) the FU Q is a slight variation of the previous question (e.g. only the wh-phrase changes, or only the object changes). This error analysis seems to suggest that it should be worthwhile to distinguish between sub-classes of topic-continuation FU Qs, and to improve specifically how answers for the "difficult" sub-classes are ranked.

The relatively high mean ranks are also due to the fact that in our approach of acquiring dialogue data, for each FU Q we marked only *one* answer from the whole repository as "correct". Again for the "complete" model, we checked the top 20 answer candidates that ranked higher than the actual "correct" one. We found that in over half of the cases an answer that could be considered correct was among the top 20.

Looking at the ranking results across the different models in Table 2, the fact that the "complete" model seems to outperform each of the three reduced models (although no statistical significance could be attained from comparing the rank numbers) confirms our two hypotheses proposed earlier. Firstly, identifying the underlying actions of questions/answers and modeling action-based sequences yield important information for identifying correct

	Reduced m. 3	**Reduced m. 2**	**Reduced m. 1**	**Complete model**
Predictors in model	tfIdfSimilarityQA	tfIdfSimilarityQA, tfIdfSimilarityAA	tfIdfSimilarityQA, lmProbQA	tfIdfSimilarityQA, tfIdfSimilarityAA, lmProbQA
Split 1	141.2	108.4	112.5	96.2
Split 2	102.7	97.4	53.8	57.7
Split 3	56.7	63.7	50.5	52.7
Split 4	40.5	26.2	37.9	35.7
Split 5	153.1	105.3	129.6	89.1
Mean	98.8	80.2	76.7	66.3

Table 2: Mean ranks of correct A_0 out of 529 answer candidates, across models and training/test splits

answers to topic continuation FU Qs. Secondly, as for the role of the immediate dialogue context for providing additional clues for identifying good answers to FU Qs, our data show that a high lexical similarity score between A_{-1} and A_0 indicates a correct answer candidate. While (Yang et al., 2006) point out the importance of Q_{-1} to provide context information, in our experiments it was generally superseded by A_{-1}.

As for the two features relating the underlying actions of A_{-1} and A_0 (actionContinuityAA, lmProbAA), the picture seems less clear; in our current modeling experiments, we had not enough evidence to keep these features. However, we plan to explore the underlying idea of action-action sequences in the future, and conjecture that such information should come into its own for *context-dependent* FU Qs.

6 Future work

Besides annotating and using more dialogue data as more people talk to our IQA system, we plan to implement a state-of-the-art topic-shift detection algorithm as proposed in (Yang et al., 2006), training and testing it on our own FU Q data. We will attempt to improve this system by adding action-based features, and then extend it to distinguish three classes: topic shifts, (topic continuation) FU Qs that are fully specified, and (topic continuation) context-dependent FU Qs. We then plan to build dedicated logistic regression models for the different sub-classes of topic continuation FU Qs. If each model uses a specific set of predictors, we hope to improve the overall rank of correct answers across the different classes of FU Qs. Also, from comparing the different models, we are interested in studying the specific properties of different FU Q types.

References

[Agresti2002] Alan Agresti. 2002. *Categorical Data Analysis*. Wiley-Interscience, New York.

[Baayen2008] R. Harald Baayen. 2008. *Analyzing Linguistic Data*. Cambridge University Press.

[Bertomeu et al.2006] Núria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 2006. Contextual phenomena and thematic relations in database QA dialogues: results from a wizard-of-oz experiment. In *Proc. of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 1–8, New York, NY.

[Dahlbäck and Jönsson1989] Nils Dahlbäck and Arne Jönsson. 1989. Empirical studies of discourse representations for natural language interfaces. In *Proc. of the 4th Conference of the European Chapter of the ACL (EACL'89)*, pages 291–298, Manchester, UK.

[Kirschner and Bernardi2007] Manuel Kirschner and Raffaella Bernardi. 2007. An empirical view on iqa follow-up questions. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

[McCoy and Cheng1991] Kathleen F. McCoy and Jeannette Cheng. 1991. Focus of attention: Constraining what can be said next. In Cecile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 103–124. Kluwer Academic Publishers, Norwell, MA.

[van Schooten et al.2009] 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.

[Yang et al.2006] 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.