Context Fusion: The Role of Discourse Structure and Centering Theory

Raffaella Bernardi, Manuel Kirschner and Zorana Ratkovic

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

Abstract

Questions are not asked in isolation. Their context, viz. the preceding interactions, might be of help to understand them and retrieve the correct answer. Previous research in Interactive Question Answering showed that context fusion has a big potential to improve the performance of answer retrieval. In this paper, we study how much context, and what elements of it, should be considered to answer Follow-Up Questions (FU Qs). Following previous research, we exploit Logistic Regression Models to learn aspects of dialogue structure relevant to answering FU Qs. We enrich existing models based on shallow features with deep features, relying on the theory of discourse structure of (Chai and Jin, 2004), and on Centering Theory, respectively. Using models trained on realistic IQA data, we show which of the various theoretically motivated features hold up against empirical evidence. We also show that, while these deep features do not outperform the shallow ones on their own, an IQA system's answer correctness increases if the shallow and deep features are combined.

1. Introduction

The goal of this paper is two-fold. First of all, we bring evidences to the importance of evaluating Language Technologies, and in particular (Interactive) Question Answering (IQA) systems, against real users' data sets. We do this by comparing TREC data against a data set of genuine human-computer dialogues. We show how the latter data significantly differ from the former. Secondly, we investigate the role of deep linguistic features to accomplish context fusion.

In (Yang et al., 2006) it is shown that shallow similarity features between a Follow-Up Question (FU Q) and the previous utterances are useful to determine whether the FU Q is a continuation of the topic of previous interaction ("topic continuation") or it is a "topic shift". The recognition of these two types of FU Qs is conjectured to be important for deciding whether or not to apply context fusion techniques for retrieving the answer. In (Kirschner et al., 2009), this conjecture has been tested by harnessing Logistic Regression Models (LRMs); the LRMs compute, for instance, whether the probability that a candidate answer to a FU Q contains the same verb of the answer provided to the previous question might help choosing the correct answer to the FU Q among all the candidate ones. The results show that in a real help-desk setting, some form of shallow context detection and fusion should be considered. In particular, the system answer preceding the FU Q seems to play an important role, especially because of its similarity to the FU Q. Both in (Chai and Jin, 2004) and (Sun and Chai, 2007) it is claimed that deeper linguistic knowledge might be necessary for deciding how much and what parts of the previous context is needed to answer a FU Q. In this paper we want to verify these claims.

In (Chai and Jin, 2004) it is stated that context question answering requires semantic-rich discourse representation structure. The authors propose a classification of possible informational transitions from one question to the other, which is meant to help deciding how context should be used in interpreting questions and retrieving answers. The proposed classification focuses on how wh-phrases, subjects and objects, verbs and their other complements vary from a question to the next. We have checked how often these types of transitions occur in a real user data set and tested whether knowing which transition has occurred in an interaction helps answering FU Qs. In (Sun and Chai, 2007), instead, entities are considered to characterize the cohesion of dialogue. Hence, the authors evaluate how dialogue models based on Centering Theory (Grosz et al., 1995; Poesio et al., 2004) succeed in processing coherent context questions, viz. topic continuation FU Q.

Our goal is, given a FU Q along with its immediately preceding utterances from the IQA dialogue, to pick the best answer from a fixed set of candidate answers, by assigning a score to each candidate, and ranking them by this score. As a theoretical result of this paper, we show which of our deep features actually describe coherence relations between utterances in IQA dialogues, thus holding up against empirical data. As a practical result of this, we show that deep features do not outperform the shallow ones discussed in (Yang et al., 2006; Kirschner et al., 2009) on their own, but do increase the answer ranking performance of an IQA system, if integrated with the latter.

2. Data sets and features

In this section, we introduce three IQA dialogue data sets, containing user-system interactions. For the purpose of calculating inter-utterance features within these user-system interactions – which we will do in Section 2.2. – 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 work 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.

2.1. Data sets

TREC'01 and TREC'04 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 a series of questions. We use two English language data sets from the 2001 and 2004 editions of the TREC QA track. The TREC'01 data set consists of 32 snippets of four turn interactions, extracted from 10 interactions, totaling 42 questions. The TREC'04 data set consists of 221 such snippets, extracted from 64 interactions, totaling 286 questions.

BoB 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 has been collected via the Bolzano Bot (BoB) that has been working as an online virtual help desk for the users of our University Library since October 2008.[1] The snippets were extracted from 916 users' interactions.

Like in TREC data, the topic continuation FU Qs can contain ellipses, e.g., Q_1: *Where can I find design books? Q_2: and dvd?*. Differently from TREC, both Q_1 and Q_2 could be just keywords, may contain noisy information such as typos or bad grammar, and could be very similar: either the user is trying to refine the question (the answer is correct but not what the user wanted to know) or the topic is further explored by moving the focus of attention to a new related entity or a new related action: 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 have been noticed also in other corpora of this type (Bertomeu, 2008; Yang et al., 2006).

2.2. Shallow and deep features

We exploit shallow features, which measure the similarity between two utterances within a snippet, and deep features, which encode the focus flow between two utterances at the task or entity level. Both for the shallow and deep features we distinguish those that relate an utterance to Q_2 (Q_2 (classification) features) and those that relate an utterance to A_2 (A_2 (identification) features). For each feature we will use names encoding the utterances involved; e.g., `A1.Q2.distsim` stands for the Distributional Similarity feature calculated between A_1 and Q_2.

Shallow features The detailed description of the shallow features 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)`: 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

- `Semantic similarity (semsim)`: Similar utterances contain similar words; we measure word-to-word similarity using WordNet (Fellbaum, 1998). We use the Lin measure, which considers also the information content of words

- `Action sequence (action)`: Based on the notion that in our help-desk setting we are dealing with task-based dialogues, which revolve around library-related actions (e.g., "borrow", "search"). Following an analysis of library-related documents, we devised a flat list of 18 library-related actions. The action feature indicates whether two utterances contain the same action. The action(s) associated with each utterance are automatically assigned by searching for strings that match words that we think represent one of our 18 actions

Deep features based on Chai and Jin's theory of discourse structure. All the deep features we present in this paper rely on information about the kinds of transitions a FU Q performs wrt. the preceding utterances in the IQA dialogue. We now turn to the transitions proposed by the theory of discourse structure of Chai and Jin (Chai and Jin, 2004). We propose the following features that implement three of their transitions:

- `Constraint Refinement (C.Ref)`: The FU Q is about a similar topic than the previous question but with additional or revised constraints

- `Participant Shift (P.Sh)`: The FU Q is about a similar topic but with different participants

- `Topic Exploration (T.Ex)`: 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) for more details):

`C.Ref`: 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

`P.Sh`: The two questions have the same syntactic predicate, but either the subject, object or argument of some preposition are different

`T.Ex`: The two questions have either the same syntactic predicate, subject, object or preposition.[2]

---

[2] We found this rather lax definition to work best in FU Q classification experiments.

We have tried other versions of these features. We tried comparing the wh-phrase of the two questions such that they are equal in the case of (a) and (b) and they are different in (c). Since the BoB data set contains many questions which do not start with a wh-phrase (see below), in another version we replaced the wh-type equivalence between the two questions with a question type equivalence, where the latter is assigned automatically (Dinh, 2009). However, the best results are achieved with no comparison of this kind, neither using the wh-type nor the question type.

Deep features based on Centering Theory. Adopting the *transition model* of (Sun and Chai, 2007), we use a four value feature, `Center.Trans`, that encodes the four transitions described below. It builds on the discourse transitions between adjacent utterances that Centering Theory introduced (Brennan et al., 1987; Grosz et al., 1995; Poesio et al., 2004). Somewhat differently from that classic theory, (Sun and Chai, 2007) defines the transitions depending on whether the head and/or the modifier of the Noun Phrases (NP) representing the *preferred centers*[3] are continued or switched between Q_1 and Q_2. Hence, the four values are:

continue: both the preferred center NP heads and NP modifiers are the same

retain: the preferred center NP heads are the same, but the NP modifiers are different

smooth shift: the preferred center NP heads are different, but the NP modifiers are the same

rough shift: both the preferred center NP heads and the NP modifiers are different

The next 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.[4] The `Center.Reference` feature evaluates to *1* (or *true*) if the noun phrase head of any antecedent is mentioned in A_2. For anaphora that are not personal pronouns, but rather definite NPs or proper names, the "antecedent" is the identical string as the anaphora.

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

---

[3]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 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 (Poesio and Kabadjov, 2004; Kabadjov, 2007) 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 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.

[4]Note that the dialogue context in this case does not include the preceding answers.

anaphora in Q_1 – using Q_2 from the previous dialogue snippet as context – the `Center.Forward` feature becomes *1* 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.

We will refer to `Center.Reference` and `Center.Forward` as A_2 deep features, since of all the deep features described in this paper, they are the ones concerning A_2 identification.

BoB vs. TREC: deep features By inspecting the corpora at disposal, we found that whereas in TREC most of the questions are wh-questions (41/42 in TREC 2001, and 279/286 in TREC 2004), in BoB data, non-wh-questions are more prevalent 2167/3044. In Table 1 we report the numbers of each type of transition, considering both Chai and Jin and Centering Theory features.[5]

Since we calculate the feature values automatically, as described in (Ratkovic, 2009), we want to assess the algorithm's precision and recall measures. For the Chai and Jin features, they are as follows: Constraint Refinement 47% recall (R) and 54% precision (P); Participant Shift: 76% R and 73% P; Topic Exploration: 81% R and 93% P.

3. Evaluation

Following (Kirschner et al., 2009), we use Logistic Regression Models (LRMs). Logistic Regression is a statistical modeling and analysis paradigm that can also be seen as a method of supervised Machine Learning. This double role makes LRM suitable for tackling both main goals of our work: modeling and analyzing the structure of IQA dialogues, and learning from dialogue data how to rank answers to FU Qs. LRMs are generalized linear models that describe the relationship between some features (independent variables) and a binary outcome (Agresti, 2002). Recall that our goal is, given a FU Q (Q_2 in our dialogue snippets), to pick the best answer from a fixed A_2 candidate set, by assigning a score to each candidate, and ranking them by this score. In our case, we have 306 answer candidates to choose from. The binary outcome of the LRM is its prediction whether each possible A_2 candidate is correct or not. The predictors in the LRM are the shallow and deep features described above. In other words, we use logistic regression to verify which of the features that have been claimed to be relevant in processing FU Q in the literature do turn out to play an important role. We will be using the following notation whenever we describe a model formula throughout this section:

`answerCorrect ~ predictor1 + predictor2`
The tilde separates the dependent variable to its left from the independent variables to its right. We try to predict whether a given A_2 is correct, considering the feature values underlying the predictors.

In all the experiments described below, we estimate the model parameters using maximum likelihood estimation. Moreover, we put each model we construct under trial by using an iterative backward elimination procedure that

---

[5]Note that we did not list counts for the Centering Theory A_2 features for TREC, since we were not able to obtain the set of (correct) A_2s for these data.

	BoB	TREC'01	TREC'04
Nr. snippets	1,522	32	221
`C.Ref`	58 (3.8%)	1 (3.1%)	18 (8.1%)
`P.Sh`	61 (4.0%)	1 (3.1%)	13 (5.9%)
`T.Ex`	428 (28.1%)	9 (28.1%)	114 (51.6%)
`center.Trans = continue`	130 (8.5%)	5 (15.6%)	89 (40.3%)
`center.Trans = retain`	24 (1.6%)	0	3 (1.4%)
`center.Trans = smoothShift`	11 (0.7%)	0	7 (3.2%)
`center.Trans = roughShift`	1357 (89.2%)	27 (84.4%)	122 (55.2%)
`center.Reference`	3%	–	–
`center.Forward`	39%	–	–

Table 1: Distribution of positive feature values in BoB and TREC data

takes off all those terms whose removal does not cause a significant drop in goodness-of-fit.[6] All the results we report in this paper are obtained with models that underwent this trimming procedure. For clarifying the modeling experiments in this section, we will present the model formulas both before and after predictor elimination.

In this work, we use LRMs to empirically verify certain theoretical claims, i.e., which particular features are informative as predictors in our models. To be able to show that our results generalize beyond our particular sample of training data, we need to validate LRMs against keeping predictors in the model that might be significant and informative only for the specific sample of training data at hand. Even if we take the mentioned measure of backward predictor elimination to eliminate uninformative predictors from our regression models, we are still in potential danger of *overfitting* the models to the training data. In fact, regression models have a tendency to overfit the training data, in that the model describes the training data well, but does not generalize well to new and unseen data.

One popular approach to validate regression models against overfitting is using *the bootstrap*. Bootstrapping is a particular method of resampling the training data, to simulate the process of sampling from the original, underlying population. We proceed as follows: for each bootstrap sample, we run the backward elimination routine described above. We then analyze which predictors were kept for how many bootstrap samples. Often enough, the resulting frequency distribution of retained predictors across the bootstrap samples sheds light on how much the set of informative predictors depends on the particular data sample: if for a majority of bootstrap samples the number of predictors that are retained by the backward elimination routine is the same, we can assume that the amount of overfitting is not problematic.

Finally, when comparing A_2 ranking results of our experimental models with their corresponding baseline models, we use two alternative hypothesis tests for checking if one model achieves better (i.e., lower) ranks for the correct A_2 than the other in a statistically significant way. Along with the mean scores of correct A_2s we will thus cite p-values,

both for the parametric paired t-test, and the non-parametric Wilcoxon signed rank test. Adopting a conservative policy, we propose to generally consider the less significant of the two tests' results for evaluating whether two particular models yield significantly different ranking results.

3.1. Baseline models used in experiments

In the first experiment, described in Section 3.2., we will test if there is empirical evidence for our deep Q_2 classification features, i.e., `C.Ref`, `P.Sh`, `T.Ex` and `center.Trans`. Each of these features is incorporated into the LRM by adding it as an *interaction term*. These interaction terms thus play the role of distinguishing between different types of FU Qs (as classified by the Q_2 features), and accordingly, to trigger different A_2 identification strategies accordingly. In our case, an interaction term provides an extra parameter to assign a differential weight to an A_2 feature depending on the value of some Q_2 feature. In the simplest case of interaction with a binary 0-1 feature (as in the case of `C.Ref`, `P.Sh`, and `T.Ex`), the interaction parameter weight is only added when the binary feature has the 1-value. We will test each Q_2 classification feature by comparing a model using that particular interaction term to a corresponding baseline model.

Baseline model 1: without interactions In this case, baseline model 1 uses a combination of the shallow A_2 identification features introduced in Section 2.2., plus all Centering Theory-based A_2 identification features. The model contains no interaction terms. Models 1 and 2 give the model specifications before and after running the backward predictor elimination procedure. We will continue using the latter, pruned model in our experiments. See Table 3 for the A_2 identification performance of this model.

```
answerCorrect ~
  action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + semsim.Q2.A2
  + center.Reference + center.Forward
```
(Model 1)

```
answerCorrect ~
  action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.Q2.A2 + semsim.A1.A2
  + center.Reference
```
(Model 2)

---

[6]Following (Harrell, 2006), we choose backward elimination, where we start with the full regression model, and keep eliminating the least significant predictors from the model, one by one, until a stopping criterion is satisfied.

Baseline model 2: with only shallow A_2 identification features As a baseline model for evaluating the effects of deep A_2 identification features, we strip baseline model 1 of its Centering Theory A_2 features. Models 3 and 4 give the model specifications before and after running the backward predictor elimination procedure; again, the latter will be used in our LRM experiments. Table 6 shows the A_2 identification performance of Model 4.

```
answerCorrect ~
  action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + semsim.Q2.A2
(Model 3)
```

```
answerCorrect ~
  action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.Q2.A2 + semsim.A1.A2 (Model 4)
```

Baseline model 3: interaction with shallow Q_2 classification feature This model takes baseline model 1 and incorporates a shallow feature as an interaction term. We pick the best-performing Q_2 classification feature from (Kirschner et al., 2009), `A1.Q2.distsim`, which we believe approximates the distinction between Topic Shift and Topic Continuity. Model specifications before and after the backward elimination routine are given in Models 5 and 6, respectively. The latter model's performance is given in Table 8.

```
answerCorrect ~ distsim.A1.Q2 *
  ( action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + semsim.Q2.A2
  + center.Reference + center.Forward )
(Model 5)
```

```
answerCorrect ~
  distsim.A1.Q2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + center.Reference
  + distsim.A1.Q2 * action.A1.A2
  + distsim.A1.Q2 * distsim.Q2.A2
(Model 6)
```

3.2. Experiments with Chai and Jin-based Q_2 classification features

Having introduced the three baseline models, we are now ready to describe the first set of modeling experiments. We add each of the three (Chai and Jin, 2004)-based features as an interaction term for Q_2 classification to baseline model 1, one at a time. In the case of `T.Ex`, the interaction term is kept in the model by the backward predictor elimination routine, which yields Model 8 from Model 7.

```
answerCorrect ~ T.Ex *
  ( action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + semsim.Q2.A2
  + center.Reference + center.Forward )
```

(Model 7)

```
answerCorrect ~
  T.Ex + action.A1.A2
  + action.Q2.A2 + lexsim.A1.A2
  + lexsim.Q2.A2 + distsim.Q2.A2
  + semsim.A1.A2 + center.Reference
  + T.Ex * action.A1.A2
  + T.Ex * lexsim.A1.A2
  + T.Ex * distsim.Q2.A2
(Model 8)
```

Model 8 also yields a minor, but nevertheless statistically significant improvement of A_2 ranking results compared to the *baseline 1* model with no interactions; Table 3 compares the ranking results, showing also how the improvement reaches statistical significance.

We perform bootstrapping to validate the model against over-fitting. Looking at the validation results provided in Table 2, and considering the selection of predictors that are retained by the backward elimination routine in the various bootstrap models, we find some variability of the number of retained predictors. After an analysis of the particular predictors that are most often dropped from the bootstrap models, we note that only the following three main effects predictors tended to get eliminated: `distsim.A1.A2`, `semsim.Q2.A2` and `center.Reference`. Some of the other factors that were occasionally dropped were interactions between `T.Ex` and some A_2 identification feature. However, *some* interaction term involving `T.Ex` generally survived the pruning procedure, which is what we are really interested in, since it shows that even if we generalize over different data samples, Q_2 classification via `T.Ex` has a *general* potential to improve a model of IQA dialogue structure.

As for the other two (Chai and Jin, 2004) features, `C.Ref` and `P.Sh`, the interaction was either dropped by the backward elimination routine, or the interactive model did not yield better A_2 ranking results than the baseline, and we do not report the model here.

3.3. Experiments with Centering Theory-based Q_2 classification feature

We now perform the same experiments on the Centering Theory-based Q_2 classification feature `center.Trans`. Model 9 shows the result of the backward elimination routine. Again, we compare the A_2 ranking performance of Model 9 to that of the main effects model *baseline 1*. Table 3 shows the minimal, yet statistically significant A_2 selection performance gain of this interactive model.

Nr. of factors retained	8	9	10	11	12	13	14	15	16
Frequency	1	2	8	28	30	19	7	3	2

Table 2: Results of 100-sample bootstrap validation of model with `T.Ex` interaction term (Model 8)

Model ID	Interaction term	Mean rank correct A_2	SD	p (Paired t-test)	p (Wilcoxon signed rank)
2 (baseline 1)	none	49.62	68.58		
8	`T.Ex`	48.95	68.35	**0.0018**	**0.0030**
9	`center.Trans`	49.12	67.96	**0.016**	**0.000006**

Table 3: Improving mean ranks of correct A_2 (out of 306 answer candidates) by adding interactions with Chai and Jin-based or Centering Theory-based features

Nr. of factors retained	9	10	11
Frequency	40	53	7

Table 4: Results of 100-sample bootstrap validation of model with `center.Trans` interaction term (Model 9)

Nr. of factors retained	7
Frequency	100

Table 5: Results of 100-sample bootstrap validation of model with added `center.Reference` predictor (Model 11)

```
answerCorrect ~
  center.Trans.num + action.A1.A2
  + action.Q2.A2 + lexsim.A1.A2
  + lexsim.Q2.A2 + distsim.Q2.A2
  + center.Trans * lexsim.A1.A2
  + center.Trans * semsim.A1.A2
  + center.Trans * center.Reference
```
(Model 9)

```
answerCorrect ~
  action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2
  + lexsim.Q2.A2 + distsim.A1.A2
  + distsim.Q2.A2 + semsim.A1.A2
  + semsim.Q2.A2 + center.Reference
  + center.Forward
```
(Model 10)

```
answerCorrect ~
  action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.Q2.A2 + semsim.A1.A2
  + center.Reference
```
(Model 11)

Finally, we validate this model against over-fitting. Table 2.2. has the validation results. Inspecting the predictors that are most likely to be dropped, we notice only `center.Reference`. We thus assume that this feature is generally less informative a feature than the other (shallow) A_2 identification features in the model. Overall, `center.Trans` seems to be more consistently informative across bootstrap samples than e.g., `T.Ex`.

3.4. Experiments with Centering Theory-based A_2 identification features

We now turn to the A_2 identification features based on Centering Theory. We test the implications of adding these "deep" A_2 identification features as predictors to a model containing the set of shallow A_2 identification features described in Section 2.2.. Models 10 and 11 show the model specifications before and after the predictor backwards elimination routine, respectively. Of the four Centering Theory-based features, only `center.Reference` is kept as a predictor in the pruned model. Table 5 shows results of checking the latter model for any signs of over-fitting: all 100 bootstrap samples retained all 7 predictors. Finally, Table 6 compares Model 11 in terms of A_2 ranking performance with the corresponding baseline model, showing how the improvement is statistically significant.

3.5. Experiments with crossed shallow and deep interaction terms

Finally, we want to explore if information from the two "deep" Q_2 classification features that we had determined to be useful (see Table 3) can further improve an interaction model that already contains a shallow feature as its interaction term. More specifically, we are interested in evidence for *three-way interactions* between a shallow and a deep Q_2 classification feature, and one of our usual shallow or deep A_2 identification features. To this end, we now introduce two models where a shallow Q_2 classification predictor is *crossed*[7] with either one of the two deep Q_2 classification features described in Table 3. Models 12 and 13 show the formulas involving `T.Ex` as the interaction term, before and after running the backward elimination routine. In the latter, one instance of a three-way interaction was deemed useful and thus retained.[8] Table 7 shows results

---

[7]The *crossed* interaction term $a \times b$ is a short-hand notation of $a + b + a * b$ in the specification of the LRM formula.

[8]We do not provide the model formulas involving the other deep feature here for presentational reasons.

Model ID	Add. feature	Mean rank correct A_2	SD	p (Paired t-test)	p (Wilcoxon signed rank)
4 (baseline 2)	-	50.35	69.00		
11	`center.Reference`	49.24	68.57	**0.00027**	**0.00003**

Table 6: Improving mean ranks of correct A_2 (out of 306 answer candidates) by adding Centering Theory-based A_2 identification features

from performing the bootstrap validation routine on Model 13. There is a wide variability in the number of predictors retained for the bootstrap models. An analysis of which predictors are often dropped from the models reveals that the three-way interaction term is very often discarded. Not surprisingly, the three-way interaction term has a high p-value in the model trained on the original data sample, signaling that there is little evidence for keeping the predictor in the model in the first place.[9] We thus have to conclude from model validation that there is not enough evidence in favor of our three-way interaction term; we attribute to over-fitting the fact that this term was actually retained in Model 13.

Table 8 compares the A_2 ranking performance of the models against a competitive baseline model with just a shallow interaction term. While the improvement caused by adding the interaction term `T.Ex` is significant only according to the non-parametric Wilcoxon test, both our statistical tests indicate significant improvements for the combination involving `center.Trans`.

```
answerCorrect ~
(distsim.A1.Q2 * T.Ex) *
  ( action.A1.A2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + semsim.Q2.A2
  + center.Reference + center.Forward )
(Model 12)
```

```
answerCorrect ~
  distsim.A1.Q2 + action.Q2.A2
  + lexsim.A1.A2 + lexsim.Q2.A2
  + distsim.A1.A2 + distsim.Q2.A2
  + semsim.A1.A2 + center.Reference
  + distsim.A1.Q2 * action.A1.A2
  + distsim.A1.Q2 * distsim.Q2.A2
  + distsim.A1.Q2 * T.Ex
              * lexsim.A1.A2
(Model 13)
```

4. Conclusion

With the A_2 ranking results in the previous section we have shown that for certain deep features based on either Chai and Jin's theory of discourse structure or on Centering Theory there is empirical evidence that they can describe the structure of realistic human-machine dialogues in the help-desk setting. Relying on the same machine-learning framework used in previous work, and building on previous results based on using shallow features to describe inter-utterance relations, we have shown that certain combinations of shallow and deep features as predictors in the models improve the models' A_2 ranking performance. A sophisticated shallow feature outperforms any of our deep features for Q_2 classification. Although we have demonstrated experimental results of how certain three-way combinations of shallow and deep features for Q_2 classification can lead to a significant improvement in our experiment, we believe that these particular findings might not hold in general for similar IQA data, but might be artifacts of over-fitting.

For the case of A_2 identification features, we have shown how features based on Centering Theory add important extra information to a model built on a powerful combination of four shallow features, again leading to a significant increase in A_2 ranking performance. In this case, the improvement is stable across different data samples.

Looking at the still relative high mean ranks in which even our best models find the correct A_2, we notice two things. Firstly, our A_2 evaluation scheme tends to be overly pessimistic, since it only considers exactly one "gold standard" answer to be correct for each given FU Q, while simply considering all remaining 305 answer candidates as completely false. However, there should clearly be major overlaps in the information content of the answer candidates, which would possibly render more than just the gold standard A_2 a good answer candidate for a particular FU Q. Secondly, for all our competitive models, the distribution of ranks of the correct answer has properties similar to those of the last model in Table 8, which we shall use as the example here. Table 9 gives descriptive statistics of the ranks of the correct A_2, out of an answer repository of 306; for half of all snippets, the correct A_2 is thus among the 12 highest-ranked candidates. The mean of the rank of correct A_2s therefore deteriorates considerably because of a rather low count of bad ranking decisions. In future work we will thus study the cases where our models tend to do worst, and thus try to find intelligent ways of improving their ranking performance.

5. References

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

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 *Pro-*

---

[9]Looking at the model statistics corresponding to the interaction term `distsim.A1.Q2 × lexsim.A1.A2 × T.Ex`, we have the following values: coefficient = 1.37, $p = 0.2525$.

Nr. of factors retained	10	11	12	13	14	15	16
Frequency	9	61	65	33	23	8	1

Table 7: Results of 200-sample bootstrap validation of model with 3-way interaction term (Model 13)

Model ID	Interaction term	Mean rank correct A_2	SD	p (Paired t-test)	p (Wilcoxon signed rank)
6 (baseline 3)	`A1.Q2.distsim`	43.95	63.91		
13	`A1.Q2.distsim` $\times$ `T.Ex`	43.71	63.98	0.0931	**0.0043**
-	`A1.Q2.distsim` $\times$ `center.Trans`	**43.15**	62.85	**0.017**	**0.0057**

Table 8: Improving mean ranks of correct A_2 (out of 306 answer candidates) with crossed, 3-way interaction terms

Summary	
Min.	1.00
1st Qu.	3.00
Median	12.00
Mean	**43.15**
3rd Qu.	58.75
Max.	305.00

Table 9: Ranks of correct A_2 out of 306 A_2 candidates

ceedings 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.

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.

Lê Thành Dinh. 2009. Question and answer classifier for closed domain interactive question answering. Master's thesis, Charles University Prague.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

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.

Frank E. Harrell. 2006. *Regression Modeling Strategies*. Springer.

Mijail Kabadjov. 2007. *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification*. Ph.D. thesis, University of Essex.

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.

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.

Massimo Poesio and Mijail Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolition module: Implementation and preliminary evaluation. In *Proc. of the 4th International Conference on Language Resources And Evaluation (LREC)*, Lisbon, Portugal.

Massimo Poesio, Rosemary Stevenson, Barbara di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

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.

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.