Evaluation of a Large-scale Email Response System

Yuval Marom and Ingrid Zukerman
Faculty of Information Technology, Monash University
Clayton, Victoria 3800, AUSTRALIA
{yuvalm,ingrid}@csse.monash.edu.au

Abstract

We are working on a large-scale, corpus-based dialogue system for responding to requests in an email-based help-desk. The size of the corpus presents interesting challenges with respect to evaluation. We discuss the limitations of the automatic evaluation performed in our previous work, and present a user study to address these limitations. We show that this user study is useful for evaluating different response generation strategies, and discuss the issue of representativeness of the sample used in the study given the large corpus on which the system is based.

1 Introduction

A help-desk domain offers interesting dialogue properties in that on the one hand responses are generalized to fit template solutions, and on the other hand they are tailored to the initiating request in order to meet specific customer needs. In recent years, we have been investigating an email-based help-desk task: generating a response to a new request based on a corpus of previous dialogues. The corpus consists of 30,000 email dialogues between customers and help-desk operators at Hewlett-Packard. However, to focus our work, we used a sub-corpus of 6,659 email dialogues which consisted of two-turn dialogues where the answers were reasonably concise (15 lines at most). These dialogues deal with a variety of customer requests, which include requests for technical assistance, inquiries about products, and queries about how to return faulty products or parts.

The size of our corpus presents challenges with respect to evaluation, which raise interesting research questions for practical corpus-based dialogue systems of a scale similar to ours. While automatic evaluations are useful during system development, the quality of a response is a subjective measure that should be judged by users of the system. Thus, user studies provide more realistic evaluations. However, how does one select a representative sample of request-response pairs to present to subjects? Many dialogue systems and other NLP systems are evaluated with user studies comprising 100-200 cases, which requires a considerable but reasonable amount of effort for test subjects and research staff. Statistically, this is an acceptable sample size when a system is based on up to 1000 cases. However, when a system is based on thousands of cases, the representativeness of such small studies is questionable. At the same time, increasing the size of a user study, and therefore the effort required from subjects and testers, may not be practical.

Is there a way to disable the NAT firewall on the CP-2W so I don't get a private ip address through the wireless network?
Unfortunately, you have reached the incorrect eResponse queue for your unit. Your device is supported at the following link, or at 888-phone-number. We apologize for the inconvenience. URL.

Figure 1: An example where terms in the request are predictive of the response.

In this paper, we report on evaluations of our email-based dialogue system, comparing different response-generation strategies. We show the limitations of an automatic evaluation of the system, and discuss a small user study that we performed in order to address these limitations. The results of our study are encouraging. However, it has its own limitations in addressing our evaluation goals. These limitations are presented as challenges for the dialogue community.

The rest of the paper is organised as follows. In the next section, we give some background to our system and its automatic evaluation. In Section 3, we present our user study, which we follow with a discussion in Section 4. In Section 5, we provide a brief review of evaluation approaches reported in similar systems, and in Section 6, we present concluding remarks.

2 Response generation strategies

2.1 Methods

In previous work we have investigated various response-generation strategies [Zukerman and Marom, 2006]. Our conclusions are that a standard retrieval approach, where a new request is matched in its entirety with previous requests or responses, is successful only in very few cases. A more suitable approach is a predictive one, which uses correlations between features of requests and responses to guide response generation [Marom and Zukerman, 2007]. Figure 1 shows an example of a request-response pair, where the terms in the

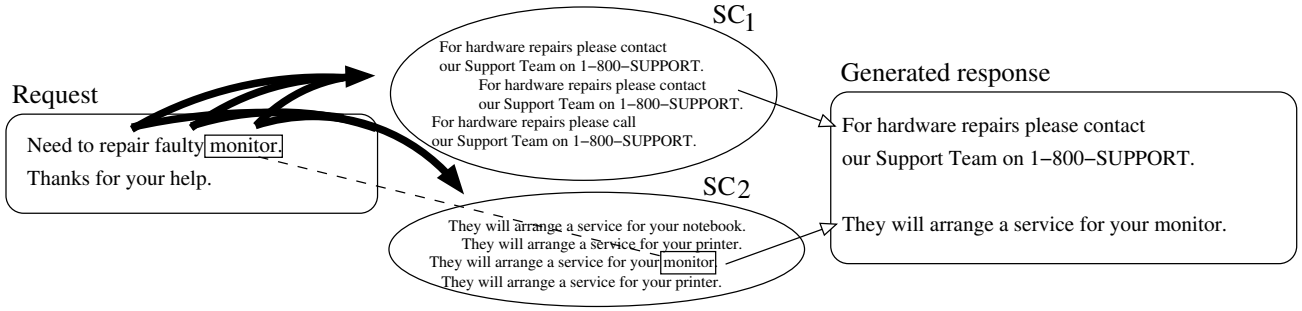

Figure 2: A fictitious example demonstrating the *Sent-Pred* and *Sent-Hybrid* methods.

response do not match any of the terms in the request, but a few of the terms in the request are predictive of the response (terms such as "firewall", "CP-2W" and "network" indicate that the query is network-related and should be redirected).

We have observed that while there is a high language variability in requests in our corpus, the responses exhibit strong regularities, mainly due to the fact that operators are equipped with in-house manuals containing prescribed answers. Further, we have observed that these regularities in the responses can occur at different levels of granularity, with two particular granularities of interest: document (email) and sentence. We have therefore implemented two predictive approaches, where we use machine learning techniques to firstly cluster responses at either of the two levels of granularity, and then learn mappings between terms in the request emails and the response clusters (document clusters or sentence clusters). We refer to these two approaches as *Document Prediction (Doc-Pred)* and *Sentence Prediction (Sent-Pred)* respectively. *Doc-Pred* produces a response by considering the response cluster with the highest prediction score, and if this prediction is higher than a confidence threshold, it selects a representative response document (the one closest to the centroid). *Sent-Pred* produces a response by considering all the sentence clusters that are predicted with sufficient confidence, and then employing multi-document summarization techniques to collate representative sentences into a single response. In the fictitious example shown in Figure 2, the combination of the terms "repair", "faulty" and "monitor" is highly predictive of sentence cluster SC_1, and the sentences in this cluster are sufficiently similar for a confident selection of a representative sentence in the response.[1] These predictive approaches only predict response clusters for which there is sufficient evidence in the corpus, and then select representative items.

In another approach, we investigated tailoring sentences to specific issues raised in the requests. We complemented the *Sent-Pred* method with a retrieval component that biases the selection of a sentence from a cluster based on how well the sentence matches any of the sentences in the request. We refer to this approach as *Sentence Prediction-Retrieval Hybrid (Sent-Hybrid)*. For example, in Figure 2, SC_2 is also highly predicted, but rather than selecting the more representative sentence (containing the term "printer"), we select the sentence that best matches the request (containing the term "monitor"). We employ this retrieval mechanism when we cannot confidently select a representative sentence from a cluster.

The two sentence-level methods (*Sent-Pred* and *Sent-Hybrid*) can produce partial responses. This happens when there is insufficient evidence to predict all the sentences required for a response. In contrast, the document-level method either produces a complete response or does not produce any response. The implementation details of these three methods are described in [Marom and Zukerman, 2007]. Here we focus on evaluation issues raised by the need to evaluate and compare these methods in the context of a very large corpus.

2.2 Automatic evaluation

In the automatic evaluation of our system we were interested in testing firstly the *coverage* of each of the methods — the proportion of requests it can address, and secondly the *quality* of the generated responses, measured separately as *correctness* and *completeness*. To measure correctness we considered the responses written by the help-desk operators as model responses, and then used the precision measure from Information Retrieval [Salton and McGill, 1983] to evaluate the response generated for each request against the model response. This measure determines the proportion of the generated response that matches the actual response. To measure completeness we used the F-Score measure, which is the harmonic mean of recall and precision (recall gives the proportion of the actual response that is included in the generated response) [Salton and McGill, 1983]. The reason for considering precision separately from the combined F-score measure is that the former simply measures whether the generated text is correct, without penalizing it for omitted information. This enables us to better assess our sentence-based methods.

The results of this evaluation are shown in Table 1, and are discussed in detail in [Marom and Zukerman, 2007]. Here we wish only to highlight a few issues. The *Doc-Pred* method produces more complete responses. This is evident from its relatively high average F-Score. Since its average precision is not higher than the precision of the other two methods, the higher F-Score must be a result of a higher average recall. However, the coverage of this method is lower than the coverage of the sentence-level methods. These methods can address additional requests, for which there is insufficient evidence for a complete response.

[1] We obtain this confidence using a measure of cluster cohesion that behaves like entropy [Marom and Zukerman, 2007].

Table 1: Results of automatic evaluation (stdev. in brackets).

Method	Coverage	Precision Ave	F-score Ave
Doc-Pred	29%	0.82 (0.21)	0.82 (0.24)
Sent-Pred	34%	0.94 (0.13)	0.78 (0.18)
Sent-Hybrid	43%	0.81 (0.29)	0.66 (0.25)

The *Sent-Pred* method produces correct responses (high precision) at the expense of completeness (low recall). The *Sent-Hybrid* method extends the *Sent-Pred* method by employing sentence retrieval as well, and thus has a higher coverage. This is because the retrieval component disambiguates between groups of candidate sentences, thus enabling more sentences to be included in a generated response. This, however, is at the expense of precision (and hence F-Score). This lower precision means that the selected sentences differ from the "selections" made by the operator in the model response. However, it does not necessarily mean that the selected sentences are worse than those used by the operator. In fact, our user-based evaluations point to situations where the opposite is the case (Section 4).

Although the automatic evaluation is valuable for comparing and fine-tuning the various methods, it has some limitations. Generated responses should be assessed on their own merit, rather than with respect to some model response, because often there is not one single appropriate response. Also, the automatic evaluation does not inform us of the usefulness of partial responses. The user study presented in the next section was designed to address these limitations.

3 User study

The aim of this study was to obtain an approximation to customers' reactions to the responses generated by the various methods, and thus provide a more subjective evaluation of our system. We asked four judges to assess the responses generated by our system. Our judges were instructed to position themselves as help-desk customers who know that they are receiving an automated response, and that such a response is likely to arrive quicker than a manual response composed by an operator.

To address the limitations of the automatic evaluation mentioned in the previous section, we designed the user study to assess the different methods from the following perspectives:

1. **Informativeness:** Is there is anything useful in the response that would make it a good automatic response, given that otherwise the customer has to wait for a human-generated response? We used a scale from 0 to 3, where 0 corresponds to "not at all informative" and 3 corresponds to "very informative".
2. **Missing information:** Are any crucial information items missing? Y/N.
3. **Misleading information:** Is there any misleading information? Y/N. We asked the judges to consider only information that might misguide the customer, and ignore information that is obviously and inconsequentially wrong, and which a customer would thus ignore, knowing that the response is automated (for example, receiving an answer for a printer, when the request was for a laptop).
4. **Compared to model response:** How does the generated response compare with the model response? Worse/Same/Better.

3.1 Experimental setup

We had two specific goals for this evaluation. First, we wanted to compare document-level versus sentence-level methods. Second, we wanted to evaluate cases where only the sentence-level methods can produce a response, and therefore establish whether such responses, which are often partial, provide a good alternative to a non-response. We therefore presented two evaluation sets to each judge.

1. The first set contained responses generated by *Doc-Pred* and *Sent-Hybrid*. These two methods obtained similar precision values in the automatic evaluation (Table 1), so we wanted to compare how they would fare with our judges.
2. The second set contained responses generated by *Sent-Pred* and *Sent-Hybrid*, for which *Doc-Pred* could not produce a response. The added benefit of this evaluation set is that it enables us to examine the individual contribution of the sentence retrieval component.

Each evaluation set contained 20 cases, randomly selected from the corpus. For each case we presented the request email, the model response email, and the two generated responses, and asked the judges to assess the generated responses on the four criteria listed above. Our four judges, who were from the Faculty of IT at Monash University, had reasonable technical experience on the kinds of issues raised in the help-desk dialogues. We asked the judges to leave a question unanswered if they felt they did not have the technical knowledge to make a judgement, but this did not actually occur.

We have chosen to maximize the coverage of this study by allocating different cases to each judge, and thus avoid a situation where a particularly good or bad set of cases is evaluated by all judges. Because the judges do not evaluate the same cases, we cannot employ standard inter-tagger agreement measures [Carletta, 1996]. However, it is nevertheless necessary to have some measure of agreement, and control for bias from specific judges or specific cases. We do this by performing pairwise significance testing, treating the data from two judges as independent samples.[2] We do this separately for each method and each of the four criteria, and then eliminate the data from a particular judge if he or she has significant disagreement with other judges. This happened with one of the judges, who was significantly more lenient than the others on the *Sent-Pred* method for the first, second and fourth criteria, and with another judge, who was significantly more stringent on the *Sent-Hybrid* method for the third criterion. Thus, each evaluation set contains a maximum of 80 cases.

[2]The statistical test employed here is the Wilcoxon rank sum test for equal medians.

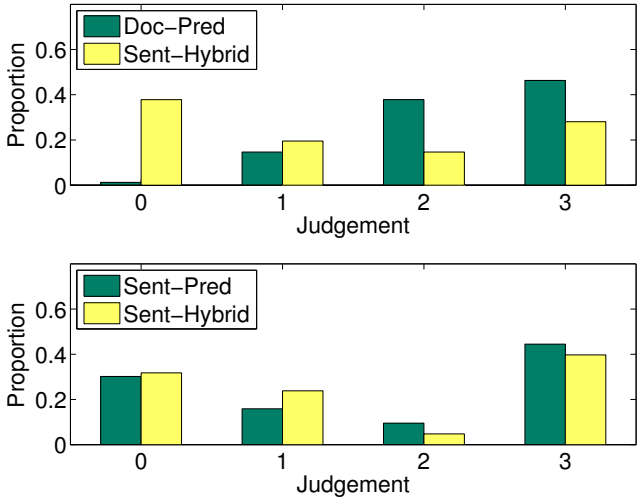

Figure 3: Evaluating the "informativeness" of generated responses.

Figure 4: Evaluating generated responses compared to model responses.

3.2 Results

Figure 3 shows the results for the "informativeness" criterion. The top part of the figure is for the first evaluation set, and it shows that when both *Doc-Pred* and *Sent-Hybrid* are applicable, the former receives an overall preference, rarely receiving a zero informativeness judgement. Since the two methods are evaluated together for the same set of cases, we can perform a paired significance test for differences between them. Using a Wilcoxon signed rank test for a zero median difference, we obtain a p-value $\ll 0.01$, indicating that the differences in judgements between the two methods are statistically significant. The bottom part of Figure 3 is for the second evaluation set, comparing the two sentence-based methods. Here there do not appear to be significant differences, and this is confirmed by the paired significance test which produces a p-value of 0.13.

Similar observations are made for the "missing information" criterion. In the first evaluation set, the *Doc-Pred* method is judged to have missing information in 23% of the cases, compared to 57% for the *Sent-Hybrid* method, and the paired significance test produces a p-value $\ll 0.01$. The second evaluation set produces a p-value of 0.11, indicating an insignificant difference between the proportions of cases judged to have missing information, which are approximately 55% for the sentence-level methods. These high proportions are in line with the low F-Scores in the automatic evaluation (Table 1): missing information results in a low recall and hence a low F-Score.

The results for the "misleading information" criterion are as follows. In the first evaluation set, 6% of the responses produced by the *Doc-Pred* method are judged to have misleading information, compared to 15% of the responses generated by the *Sent-Hybrid* method. Although the proportion of misleading information is higher for the latter, the paired differences between the two methods are not statistically significant, with a p-value 0.125. For the second evaluation set, the proportions are 11% and 10% for *Sent-Pred* and *Sent-Hybrid* respectively, and their paired differences are also insignificant with a p-value of 1.0. These low proportions of misleading information are in line with the high precision values observed from the automatic evaluation (Table 1): while responses with a high precision may be incomplete, they generally contain correct information.

Lastly, the results for the "compared to model response" criterion are shown in Figure 4. The top part of the figure, corresponding to the first evaluation set, shows that *Doc-Pred* receives more "same" than "worse" judgements, compared to *Sent-Hybrid*, and they both receive a small proportion of "better" judgements. The paired significance test produces a p-value $\ll 0.01$, confirming that these differences are significant. The bottom part of the figure, corresponding to the second evaluation set, shows smaller differences between *Sent-Pred* and *Sent-Hybrid*, and indeed the p-value for the paired differences is 0.27. Notice that *Sent-Pred* does not receive any "better" judgements, while *Sent-Hybrid* does.

4 Discussion

The results from our previous work (Table 1) showed that the different response-generation strategies are all able to address a significant proportion of the requests, with varying degrees of success. These results were obtained through an automatic evaluation that performed a textual comparison between a generated response and the actual response for a given request. However, these results only provided a limited insight into whether the different strategies achieved their aims. The user study presented in this paper enabled us to evaluate specific characteristics that could only be judged subjectively.

- ***Doc-Pred.*** This document-level strategy attempts to reuse a complete response in the corpus for a new request. The results show that when such a strategy is possible it is better than a sentence-level strategy: the generated response is more informative and complete, and compares more favourably with the model response.
- ***Sent-Pred.*** This sentence-level strategy attempts to produce a response with as much content as is warranted by

evidence in the corpus. Hence, this strategy can yield a partial response. Our results show that indeed this strategy can miss crucial information, but what it does include in a response can be informative and is rarely misleading. The responses are sometimes as good as the model responses.

- ***Sent-Hybrid.*** This hybrid strategy is based on the *Sent-Pred* strategy, but it attempts to tailor a response to specific terms in the request. The main difference between this strategy and *Sent-Pred* is that rather than selecting a representative sentence from a sentence cluster, it selects a sentence that best matches the request. Hence, the generated responses are less general than those produced by *Sent-Pred* (and sometimes less general than the model responses). However, there were no statistically significant differences between the two strategies on any of the criteria we measured in the user study.

It is encouraging that the performance of *Sent-Hybrid* is at least as good as that of *Sent-Pred*, because we saw in the automatic evaluation that *Sent-Hybrid* has a higher coverage (Table 1). However, it is somewhat surprising that *Sent-Hybrid* did not outperform *Sent-Pred* overall. It is worth noting that in a few of the cases, *Sent-Hybrid* produced a better response than the model response. That is, the judges thought that the generated response contained additional useful information not appearing in the model response. However, this did not occur sufficiently to show up significantly in the results.

The similar performance of the two sentence-level methods may be due to a genuine insignificant effect from the retrieval component of the *Sent-Hybrid* method, or due to the fact that an effect could not be observed in the sample that was selected for the user study. Therefore, although the user study was valuable in showing that sentence-level strategies provide useful alternatives when document-level ones cannot be used, it was limited in that it left an aspect of our research inconclusive.

The data in the user study account for 2.4% of the corpus used in the automatic evaluation. Our corpus is divided into topic-based datasets. The data for the user study were selected from these different datasets in proportion to the number of dialogues in each topic. Although this data-selection policy makes the test set fair, it increases the difficulty of drawing specific conclusions. For example, it would be difficult to determine whether a particular response-generation strategy is more suitable for specific topics or for particular kinds of requests. In order to test such possibilities we would need to increase the sample size substantially. Alternatively, we could conduct preliminary automated evaluations for specific conditions, and then target these conditions in user-based evaluations.

These observations point to the need to balance the requirements derived from large corpora with the affordances provided by human subjects. That is, as the sizes of dialogue corpora increase, and several operating parameters of a system need to be considered, the number of requisite trials increases as well. At the same time, the amount of data that subjects can evaluate is limited, more so when they are required to read and judge long texts. These issues must be considered in tandem to devise appropriate sampling protocols for user studies.

Although large-scale, corpus-based systems are being routinely evaluated automatically in NL systems, scant attention has been given to the determination of a suitable sample size for trials with people (Section 5). In contrast, human experiments conducted in the social and medical sciences are concerned with sample sizes. Sampling methodologies, such as power analysis, were developed to help experimenters plan sample sizes [Lewicki and Hill, 2006]. These methodologies take into account factors such as measurement error, type of statistical test used, and desired level of significance. Although some of these factors, such as measurement error, are not always relevant for NL and dialogue systems, we can nevertheless use these methodologies in our studies.

5 Evaluation in related research

A comprehensive review of existing evaluation methods for practical dialogue systems is outside the scope of this paper. Instead, we mention a few systems we have encountered recently, with a particular emphasis on dialogue systems, whose response strategies rely on a corpus, and hence the usefulness of the evaluation depends on the size of the corpus.

There are two systems where the corpus is similar in size to our corpus [Berger and Mittal, 2000; Carmel *et al.*, 2000]. The corpus of the system described in [Berger and Mittal, 2000] consisted of 10,395 call-center request-response pairs, of which they used 70% for training, and 30% for testing. The evaluation on the test set, which is automatic, examined the rank of the real response in the retrieved list. Like our automatic evaluation, this approach assumed that the real response is the best one, but unlike our automatic evaluation, there was no consideration of other responses that might be similar to the real response. That is, the responses that appear near the real response in the retrieved list were not evaluated. The eResponder system [Carmel *et al.*, 2000] retrieved a list of request-response pairs from a corpus and presented a ranked list of responses for a given query. The corpus was an NSF archive called "Ask a Scientist or Engineer", whose size is not mentioned in the paper, but an internet report states that it has "thousands of questions" (`http://content.nsdl.org/wbr/Issue.php?issue=44`). The system was evaluated on 30 queries by a user-based evaluation of the relevance of the top 3, 5 and 10 retrieved responses.

Both of these systems returned a list of responses to the user — they did not attempt to produce a single response. This means that they are concerned with different evaluation issues than those considered here.

Four examples of systems which use smaller corpora are reported in [Lapalme and Kosseim, 2003; Roy and Subramaniam, 2006; Feng *et al.*, 2006; Leuski *et al.*, 2006]. Lapalme and Kosseim's system involved a corpus of 1,568 email dialogues, and evaluations of two tasks: a classification task tested on 144 annotated cases, and a retrieval task tested on 102 cases. Roy and Subramaniam's system involved a corpus of 2,000 transcribed call-center calls, and an evaluation of a clustering task tested on 125 annotated cases. In both of these systems there is also a response-generation task, which was

not evaluated. In contrast, Feng *et al.* evaluated their response generation module. Their corpus consisted of 1,236 discussion threads, and they performed a manual evaluation where judges used a criterion similar to our "informativeness" criterion to assess the quality of responses as "exact answer", "good answer", "related answer" or "unrelated answer". The test set contained 66 cases, which is 5.4% of the corpus. This is almost double the proportion of our user study, but the size of the corpus is less than a fifth of ours. Finally, Leuski *et al.*'s system was based on a corpus of 1,261 questions. However, its animated character can utter only 83 distinct responses to these questions. Their test set consisted of 20 subjects, where each asked the system 20 questions. A manual evaluation was then carried out by three judges, who rated the system's responses on a 6-scale criterion that takes into account both utility and discourse quality.

To summarize, the above systems illustrate different types of response-generation tasks, which are evaluated using a variety of criteria, both in automatic and user-based evaluations. However, in the latter, the representativeness of the evaluation sets was not considered. With the increased availability of electronic resources, the consideration of these issues is timely for the dialogue and NL communities.

6 Conclusion

In this paper, we have discussed the evaluation of our corpus-based, response-generation strategies for an email-based, help-desk dialogue system. The various strategies take advantage of the strong regularities that exist in help-desk responses, by abstracting them either at the document level or at the sentence level. They then find correlations between requests and responses to build predictive models for addressing new requests. The hybrid method we presented was designed to overcome the loss of information resulting from abstracting response sentences. The deployment of sentence retrieval in combination with prediction was shown to be useful for better tailoring a response to a request. Our results show that each of the strategies can address a significant portion of the requests, and that when the re-use of a complete response is not possible, the collation of sentences into a partial response can be useful.

We identified limitations of our automatic evaluation, and presented a user study where human judgements provide a more subjective indication of the quality of the generated responses. Although this study addressed some of the limitations of the automatic evaluation, it also posed questions regarding the sampling of data for user-based studies of dialogue systems driven by a large corpus. We have not seen many dialogue systems of this kind in the literature. However, with the constant increase of digital information and archives, more of these systems will be developed, necessitating answers to the questions we have raised.

Acknowledgments

This research was supported in part by grant LP0347470 from the Australian Research Council and by an endowment from Hewlett-Packard. The authors also thank Hewlett-Packard for the extensive anonymized help-desk data.

References

[Berger and Mittal, 2000] A. Berger and V.O. Mittal. Query-relevant summarization using FAQs. In *ACL2000 – Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 294–301, Hong Kong, 2000.

[Carletta, 1996] J. Carletta. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[Carmel *et al.*, 2000] D. Carmel, M. Shtalhaim, and A. Soffer. eResponder: Electronic question responder. In *CoopIS '02: Proceedings of the 7th International Conference on Cooperative Information Systems*, pages 150–161, Eilat, Israel, 2000.

[Feng *et al.*, 2006] D. Feng, E. Shaw, J. Kim, and E. Hovy. An intelligent discussion-bot for answering student queries in threaded discussions. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 171–177, Sydney, Australia, 2006.

[Lapalme and Kosseim, 2003] G. Lapalme and L. Kosseim. Mercure: Towards an automatic e-mail follow-up system. *IEEE Computational Intelligence Bulletin*, 2(1):14–18, 2003.

[Leuski *et al.*, 2006] A. Leuski, R. Patel, D. Traum, and B. Kennedy. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia, 2006.

[Lewicki and Hill, 2006] P. Lewicki and T. Hill. *Statistics: Methods and Applications*. StatSoft Inc, Tulsa, Oklahoma, 2006.

[Marom and Zukerman, 2007] Y. Marom and I. Zukerman. A predictive approach to help-desk response generation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India, 2007.

[Roy and Subramaniam, 2006] S. Roy and L.V. Subramaniam. Automatic generation of domain models for call-centers from noisy transcriptions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 737–744, Sydney, Australia, 2006.

[Salton and McGill, 1983] G. Salton and M.J. McGill. *An Introduction to Modern Information Retrieval*. McGraw Hill, 1983.

[Zukerman and Marom, 2006] I. Zukerman and Y. Marom. A comparative study of information-gathering approaches for answering help-desk email inquiries. In *Proceedings of the 19th ACS Australian Joint Conference on Artificial Intelligence*, Hobart, Australia, 2006.