A predictive approach to help-desk response generation

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

Abstract

We are developing a corpus-based approach for the prediction of help-desk responses from features in customers' emails, where responses are represented at two levels of granularity: document and sentence. We present an automatic and human-based evaluation of our system's responses. The automatic evaluation involves textual comparisons between generated responses and responses composed by the help-desk operators. The results show that both levels of granularities produce good responses, addressing inquiries of different kinds. The human-based evaluation measures response informativeness, and confirms our conclusion that both levels of granularity produce useful responses.

1 Introduction

Email inquiries sent to help desks often "revolve around a small set of common questions and issues".[1] This means that help-desk operators spend most of their time dealing with problems that have been previously addressed. Further, a significant proportion of help-desk responses contain a low level of technical content, corresponding, for example, to inquiries addressed to the wrong group, or insufficient detail provided by the customer. Organizations and clients would benefit if the efforts of human operators were focused on difficult, atypical problems, and an automated process was employed to deal with the easier problems.

In this paper, we report on our experiments with corpus-based approaches for the automation of help-desk responses. Our study was based on a log of 30,000 email dialogues between users and help-desk operators at Hewlett-Packard. However, to focus our work, we used a sub-corpus of 6659 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 user requests, which include requests for technical assistance, inquiries about products, and queries about how to return faulty products or parts.

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.

Figure 1: A sample request-response pair.

Analysis of our corpus reveals that requests containing precise information, such as product names or part specifications, sometimes elicit helpful, precise answers referring to this information, while other times they elicit answers that do not refer to the query terms, but contain generic information, as seen in the example in Figure 1.[2] Our previous experiments show that a standard document retrieval approach, where a new request is matched in its entirety with previous requests or responses, is only successful in very few cases [Zukerman and Marom, 2006]. We posit that this is because (1) many requests raise multiple issues, and hence do not match well any one document; (2) the language variability in the requests is very high; and (3) as seen in Figure 1, the replies to many technical requests are largely non-technical, and hence do not match technical terms in the requests.

These observations lead us to consider a predictive approach, which uses correlations between features of requests and responses to guide response generation. In principle, correlations could be modelled directly between terms in the requests and responses [Berger and Mittal, 2000]. However, we have observed that responses in our corpus exhibit strong regularities, mainly due to the fact that operators are equipped with in-house manuals containing prescribed answers. For example, Figure 2 shows two responses that contain parts that are almost identical (the two italicized sentences). The existence of these regularities motivates us to generate abstractions of responses, rather than deal with low-level response terms. In contrast, we choose to represent requests at a low level, due to the high language variability mentioned above. The request-response correlations are then modeled at these two levels. As seen in the examples in Figures 1 and 2, the desirable granularity for representing responses can vary: it can be as fine as sentences (Figure 2), or as coarse as complete documents (Figure 1). The investigations reported here involve these two levels of granularity, leading to the *Sent-Pred* and *Doc-Pred* methods respectively (Section 2).

[1] http://customercare.telephonyonline.com/ar/telecom_next_generation_customer.

[2] Sample examples are reproduced verbatim from the corpus (except for URLs and phone numbers, which have been disguised by us), and some have customer or operator errors.

> If you are able to see the Internet then it sounds like it is working, you may want to get in touch with your IT department to see if you need to make any changes to your settings to get it to work. *Try performing a soft reset, by pressing the stylus pen in the small hole on the bottom left hand side of the Ipaq and then release.*

> *I would recommend doing a soft reset by pressing the stylus pen in the small hole on the left hand side of the Ipaq and then release*. Then charge the unit overnight to make sure it has been long enough and then see what happens. If the battery is not charging then the unit will need to be sent in for repair.

Figure 2: Responses that share a sentence.

As for any learning task, building prediction models for responses at an abstracted level of representation has advantages and drawbacks. The advantages are that the learning is more focused, and it deals with data of reduced sparsity. The drawback is that there is some loss of information, when somewhat dissimilar response units (sentences or documents) are grouped together. In order to overcome this disadvantage, we have developed a prediction-retrieval hybrid approach, which predicts groups of responses, and then selects between dissimilar response units by matching them with request terms. We investigate this approach at the sentence level, leading to the *Sent-Hybrid* method (Section 2).

Note that the *Doc-Pred* method essentially re-uses an existing response in the corpus to address a new request. In contrast, the two sentence-based methods combine sentences from multiple response documents to produce a new response, as is done in multi-document summarization [Filatova and Hatzivassiloglou, 2004]. Hence, unlike *Doc-Pred*, *Sent-Pred* and *Sent-Hybrid* may produce partial responses. In this paper, we investigate when the different methods are applicable, and whether individual methods are uniquely successful in certain situations. Specifically, we consider the trade off between partial, high-precision responses, and complete responses that may contain irrelevant information.

The rest of this paper is organized as follows. In the next section, we describe our three prediction methods, followed by the evaluation of their results in Section 3. In Section 4, we discuss related research, and then present our conclusions and plans for future work in Section 5.

2 Methods

In this section, we present the implementation details of our three prediction methods. Note that some of the methods were implemented independently of each other at different stages of our project. Hence, there are minor implementational variations, such as choice of machine learning algorithms and some discrepancies regarding features. We plan to bridge over these differences in the near future, but are confident that they do not impede the aim of the current study: evaluating a predictive approach to response generation.

2.1 Document Prediction (*Doc-Pred*)

This prediction method first groups similar response documents (emails) in the corpus into response clusters. For each request it then predicts a response cluster on the basis of the request features, and selects the response that is most representative of the cluster (closest to the centroid). This method would predict a group of responses similar to the response in Figure 1 from the input term “CP-2W”.

The clustering is performed in advance of the prediction process by the clustering program *Snob*, which performs mixture modelling combined with model selection based on the Minimum Message Length criterion [Wallace and Boulton, 1968]. We chose this program because one does not have to specify in advance the number of clusters. We use a binary representation whereby the lemmatized content words in the corpus make up the components of an input vector, and its values correspond to the absence or presence of each word in the response (this representation is known as bag-of-words).

The predictive model is a Decision Graph [Oliver, 1993] trained on (1) input features: unigram and bigram lemmas in the request,[3] and (2) target feature: the identifier of the response cluster that contains the actual response for the request. The model provides a prediction of which response cluster is most suitable for a given request, as well as a level of confidence in this prediction. If the confidence is not sufficiently high, we do not attempt to produce a response.

2.2 Sentence Prediction (*Sent-Pred*)

As for the *Doc-Pred* method, the *Sent-Pred* method starts by abstracting the responses. It uses the same clustering program, *Snob*, to cluster sentences into *Sentence Clusters (SCs)*, using the representation used for *Doc-Pred*.[4] Unlike the *Doc-Pred* method, where only a single response cluster is predicted, resulting in a single response document being selected, in the *Sent-Pred* method several SCs are predicted. This is because we are trying here to collate multiple sentences into one response. Each request is used to predict promising SCs, and a response is composed by extracting one sentence from each such SC. Because the sentences in each SC originate from different response documents, the process of selecting them for a new response corresponds to multi-document summarization. In fact, our selection mechanism is based on a multi-document summarization formulation proposed by Filatova and Hatzivassiloglou [2004].

To illustrate these ideas, consider the fictitious example in Figure 3. Three small SCs are shown in the example (in practice the SCs can have tens and hundreds of sentences). The thick arrows correspond to high-confidence predictions, while the thin arrows correspond to sentence selection. The other components of the diagram demonstrate the *Sent-Hybrid* approach (Section 2.3). In this example, three of the request terms – “repair”, “faulty” and “monitor” – result in a confident prediction of two SCs: SC_1 and SC_2. The sentences in SC_1 are identical, so we can arbitrarily select a sentence to include in the generated response. In contrast, although the sentences in SC_2 are rather similar, we are less confident in arbitrarily selecting a sentence from it.

The predictive model is a Support Vector Machine (SVM). A separate SVM is trained for each SC, with unigram and bi-

[3] Significant bigrams are obtained using the NSP package (`http://www.d.umn.edu/˜tpederse/nsp.html`).

[4] We have also experimented with syntactic features of sentences, but the simple bag-of-words representation was as competitive.

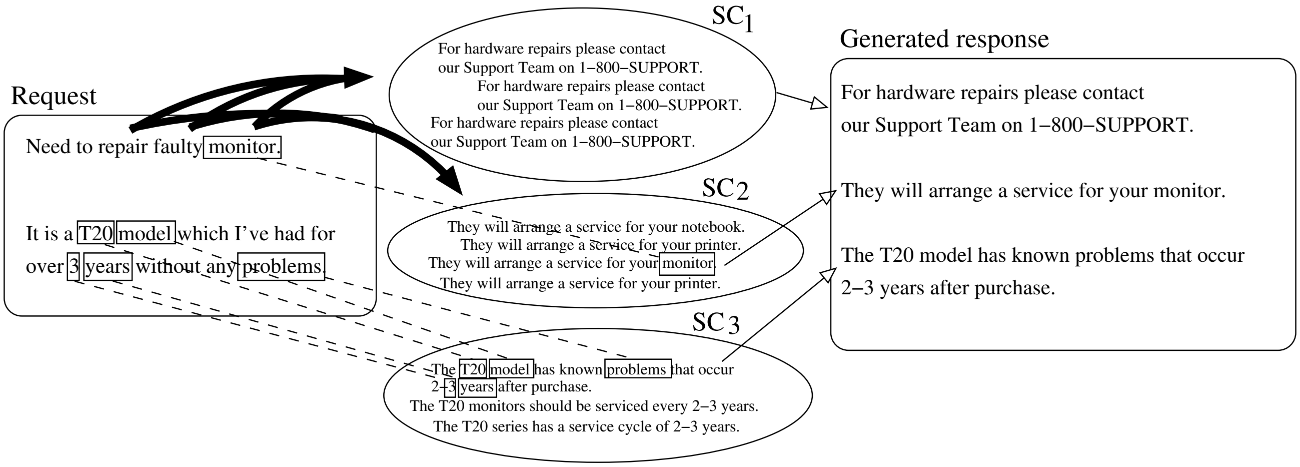
Request
Need to repair faulty monitor.
It is a T20 model which I've had for over 3 years without any problems.
SC1
For hardware repairs please contact our Support Team on 1−800−SUPPORT.
For hardware repairs please contact our Support Team on 1−800−SUPPORT.
For hardware repairs please contact our Support Team on 1−800−SUPPORT.
SC2
They will arrange a service for your notebook.
They will arrange a service for your printer.
They will arrange a service for your monitor
They will arrange a service for your printer.
SC3
The T20 model has known problems that occur 2−3 years after purchase.
The T20 monitors should be serviced every 2−3 years.
The T20 series has a service cycle of 2−3 years.
Generated response
For hardware repairs please contact our Support Team on 1−800−SUPPORT.
They will arrange a service for your monitor.
The T20 model has known problems that occur 2−3 years after purchase.

Figure 3: A fictitious example demonstrating the *Sent-Pred* and *Sent-Hybrid* methods.

gram lemmas in a request as input features, and a binary target feature specifying whether the SC contains a sentence from the response to this request. During the prediction stage, the SVMs predict zero or more SCs for each request, as shown in Figure 3. The sentence closest to the centroid is then extracted from each highly *cohesive* SC predicted with *high confidence*. A high-confidence prediction indicates that the sentence is relevant to many requests that share certain regularities. A cluster is cohesive if the sentences in it are similar to each other, which means that it is possible to obtain a sentence that represents the cluster adequately. These stringent requirements placed on confidence and cohesion mean that the *Sent-Pred* method often yields partial responses.

The cohesion of an SC is calculated as

$$\frac{1}{N}\sum_{k=1}^{N}[\Pr(w_k \in SC) \le \alpha \ \vee \ \Pr(w_k \in SC) \ge 1-\alpha]$$

where N is the number of content lemmas, $\Pr(w_k \in SC)$ is the probability that lemma w_k is used in the SC (obtained from the centroid), and α is a parameter that represents how strict we are when judging the similarity between sentences. We have used this parameter, instead of raw probabilities, because strictness in judging sentence similarity is a subjective matter that should be decided by the users of the system. The sensitivity analysis we performed for this parameter is discussed in [Marom and Zukerman, 2005].

Our formula implements the idea that a cohesive group of sentences should agree on both the words that are included in these sentences and the words that are omitted. For instance, the italicized sentences in Figure 2 belong to an SC with cohesion 0.93. For values of α close to zero, the equation above behaves like entropy, favouring very strong agreement on word usage and omission. The latter is necessary because just knowing that certain words appear in a high percentage of the sentences in a cluster is not sufficient to determine how similar are these sentences. One also must know that these sentences exclude most other words.

2.3 Sentence Prediction-Retrieval (*Sent-Hybrid*)

As we can see in cluster SC_2 in Figure 3, it is possible for an SC to be strongly predicted without being sufficiently cohesive for a confident selection of a representative sentence. However, sometimes the ambiguity can be resolved through cues in the request. In this example, one of the sentences matches the request terms better than the other sentences, as it contains the word "monitor". The *Sent-Hybrid* method complements the prediction component of *Sent-Pred* with a retrieval component, and thus forms a hybrid.

This retrieval component implements the traditional Information Retrieval paradigm [Salton and McGill, 1983], where a "query" is represented by its content terms, and the system retrieves a set of documents that best matches this query. In the help-desk domain, a good candidate for sentence retrieval contains terms that appear in the request, but also contains other terms that hopefully provide the requested information (in the example in Figure 3, the "best" sentence in SC_2 shares only one term with the request). Therefore, we perform a recall-based retrieval, where we find sentences that match as many terms as possible in the request, but are allowed to contain additional terms. Recall is calculated as

$$\textit{recall} = \frac{\Sigma \ \text{TF.IDF of lemmas in request sent \& response sent}}{\Sigma \ \text{TF.IDF of lemmas in request sentence}}$$

We have decided to treat the individual sentences in a request email as separate "queries", rather than treat the complete email as a single query, because a response sentence is more likely to have a high recall when matched against a single request sentence as opposed to a whole document.

For highly cohesive SCs predicted with high confidence, we select a representative sentence as before.

For SCs with medium cohesion predicted with high confidence, we attempt to match its sentences with a request sentence. Here we use a liberal (low) recall threshold, because the high prediction confidence guarantees that the sentences in the cluster are suitable for the request. The role of retrieval in this situation is to select the sentence whose content lemmas best match the request, regardless of how well they match (the non-content lemmas, also known as function words or stop words, are excluded from this account).

For uncohesive clusters or clusters predicted with low confidence, we can rely only on retrieval. Now we must use a more conservative recall threshold to ensure that only very highly-matching sentences are included in the response. SC_3 in Figure 3 is an example of an SC for which there is insufficient evidence to form strong correlations between it and request terms. However, we can see that one of its sentences matches very well the second sentence in the request. In fact, all the content words in that request sentence are matched, resulting in a perfect recall score of 1.0.

Once we have the set of candidate response sentences that satisfy the appropriate recall thresholds, we remove redundant sentences. Since sentences that belong to the same medium-cohesive SC are quite similar to each other, it is sufficient to select a single sentence – that with the highest recall. Sentences from uncohesive clusters are deemed sufficiently different to each other, so they can all be retained. All the retained sentences will appear in the generated response (at present, these sentences are treated as a set, and are not organized into a coherent reply).

3 Evaluation

In this section, we examine the performance of our three prediction methods. Our experimental setup involves a standard 10-fold validation, where we repeatedly train on 90% of a dataset and test on the remaining 10%. We present each test split as the set of new cases to be addressed by the system.

3.1 Measures

We are interested in two performance indicators: *coverage* and *quality*.

Coverage is the proportion of requests for which a response can be generated. We wish to determine whether the three methods presented in the previous section are applicable in different situations, i.e., how exclusively they address or "cover" different requests. Each of these methods specifies a requisite level of confidence for the generation of a response. If the response planned by a method fails to meet this confidence level for a request, then the request is not covered by this method. Since the sentence-based methods generate partial responses, we say that their responses cover a request if they contain at least one sentence generated with high confidence. Non-informative sentences, such as "Thank you for contacting HP", which are often produced by these methods, have been excluded from our calculations, in order to have a useful comparison between our methods.

Quality is a subjective measure, best judged by users of a deployed system. Here we approximate a quality assessment by means of two experiments: a preliminary human-based study where people evaluate a small subset of the responses generated by our system (Section 3.3); and a comprehensive automatic evaluation that treats the responses generated by the help-desk operators as model responses, and performs text-based comparisons between these responses and the generated ones (Section 3.2). We are interested in the *correctness* and *completeness* of a generated response. The former measures how much of its information is correct, and the latter measures its overall similarity with the model response. We consider correctness separately because it does not penalize missing information, enabling us to better assess our sentence-based methods. These measures are approximated by means of two measures from Information Retrieval [Salton and McGill, 1983]: *precision* approximates correctness, and *F-score* approximates completeness and correctness in combination. Precision and F-score are calculated as follows using a word-by-word comparison (stop-words are excluded).[5]

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

Method	Coverage	Quality	
		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)

$$\text{Precision} = \frac{\text{\# words in both model and generated response}}{\text{\# of words in generated response}}$$

$$\text{Recall} = \frac{\text{\# words in both model and generated response}}{\text{\# of words in model response}}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.2 Results

The combined coverage of the three methods is 48%. This means that for 48% of the requests, a partial or complete response can be generated by at least one of the methods. Table 1 shows the individual results obtained by the different methods. The *Doc-Pred* method can address 29% of the requests. Only an overall 4% of the requests are uniquely addressed by this method, but the generated responses are of a fairly high quality, with an average precision and F-score of 0.82. Notice the large standard deviation of these averages, suggesting a somewhat inconsistent behaviour. This is due to the fact that this method gives good results only when complete generic responses are found. In this case, any re-used response will have a high similarity to the model response. However, when this is not the case, the performance degrades substantially, resulting in inconsistent behaviour. This means that *Doc-Pred* is suitable when requests that share some regularity receive a complete template response.

The *Sent-Pred* method can find regularities at the sub-document level, and therefore deal with cases where partial responses can be generated. It produces responses for 34% of the requests, and does so with a consistently high precision (average 0.94, standard deviation 0.13). Only an overall 1.6% of the requests are uniquely addressed by this method, however, for the cases that are shared between this method and other ones, it is useful to compare the actual quality of the generated responses. For example, with respect to *Doc-Pred*, there are 10.5% cases where *Sent-Pred* either uniquely addresses requests, or jointly addresses requests but has a higher F-score. This means that in some cases a partial response has a higher quality than a complete one. Note that since *Sent-Pred* has a higher average precision than *Doc-Pred*, its lower average F-Score must be due to a lower average recall. This confirms that *Sent-Pred* produces partial responses.

[5] We have also employed sequence-based measures using the ROUGE tool set [Lin and Hovy, 2003], with similar results to those obtained with the word-by-word measures.

My screen is coming up reversed (mirrored). There must be something loose electronically because if I put the stylus in it's hole and move it back and forth, I can get the screen to display properly momentarily. Please advise where to send for repairs.

To get the iPAQ serviced, you can call `1-800-phone-number`, options 3, 1 (enter a 10 digit phone number), 2. Enter your phone number twice and then wait for the routing center to put you through to a technician with Technical Support. They can get the unit picked up and brought to our service center.

To get the iPAQ repaired (battery, stylus lock and screen), please call `1-800-phone-number`, options 3, 1 (enter a 10 digit phone number), 2.

Figure 4: Example demonstrating the *Sent-Hybrid* method.

The *Sent-Hybrid* method extends the *Sent-Pred* method by employing sentence retrieval, and thus has a higher coverage (43%). This is because the retrieval component can often include sentences from SCs with medium and low cohesion, which might otherwise be excluded. However, this is at the expense of precision. Retrieval selects sentences that match closely a given request, but this selection can differ from the "selections" made by the operator in the model response. Precision (and hence F-score) penalizes such sentences, even when they are more appropriate than those in the model response. For example, consider the request-response pair at the top of Figure 4. The response is quite generic, and is used almost identically for several other requests. The *Sent-Hybrid* method almost reproduces this response, replacing the first sentence with the one shown at the bottom of Figure 4. This sentence, which matches more request words than the first sentence in the model response, was selected from an SC that is not highly cohesive, and contains sentences that describe different reasons for setting up a repair (the matching word is "screen"). Overall, the *Sent-Hybrid* method outperforms the other methods in about 12% of the cases, where it either uniquely addresses requests, or addresses them jointly with other methods but produces responses with a higher F-score.

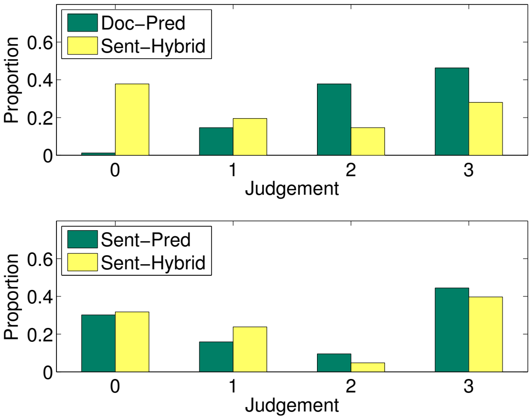

Figure 5: Human judgements of informativeness.

3.3 Human judgements

The purpose of this part of the evaluation is twofold. First, we want to compare the quality of responses generated at different levels of granularity. Second, we want to evaluate cases where only the sentence-based methods can produce a response, and therefore establish whether such responses, which are often partial, provide a good alternative to a non-response. Hence, we constructed two evaluation sets: one containing responses generated by *Doc-Pred* and *Sent-Hybrid*, and one containing responses generated by *Sent-Pred* and *Sent-Hybrid*. The latter evaluation set enables us to further examine the contribution of the retrieval component of the hybrid approach. Each evaluation set comprises 20 cases, and each case contains the request email, the model response email, and the two system-generated responses. We asked four judges to rate the generated responses on several criteria. Owing to space limitations, we report here only on one of these: informativeness. We used a scale from 0 to 3, where 0 corresponds to "not at all informative" and 3 corresponds to "very informative". The judges were instructed to position themselves as users of the system, who know that they are receiving an automated response, which is likely to arrive faster than a response composed by an operator.

We maximized the coverage of this study by allocating different cases to each judge, thus avoiding a situation where a particularly good or bad set of cases is evaluated by all the judges. Since the judges do not evaluate the same cases, we cannot employ standard inter-tagger agreement measures. Still, it is necessary to have some measure of agreement between judges, and control for bias from specific judges or specific cases. We do this separately for each prediction method by performing pairwise significance testing (using the Wilcoxon rank sum test for equal medians), where the data from two judges are treated as independent samples. We then eliminate the data from a particular judge if he or she has significant disagreement with other judges. This happened with one judge, who was significantly more lenient than the others on the *Sent-Pred* method. Since there are four judges, we have an overall maximum of 80 cases in each evaluation set.

Figure 5 shows the results for the two evaluation sets. The top part, which is for the first set, 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 5 is for the second evaluation set, comparing the two sentence-based methods. Recall that this evaluation set comprises cases that were only addressed by these two methods. Thus, the first important observation from this chart is that when a complete response cannot be re-used, a response collated from individual sentences is often judged to contain some level of informativeness. The second observation from this chart is that there does not seem to be a difference between the two methods. In fact, the above paired significance test produces p-value of 0.13 for the second evaluation set, thus confirming that the differences are not statistically significant. 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. However, it is somewhat surprising that *Sent-Hybrid* did not perform better than *Sent-Pred*. It is worth noting that there were a few cases where judges commented that a generated response contained additional useful information not appearing in the model response, as seen in the example in Figure 4. We confirmed that these responses were generated by the *Sent-Hybrid* method, but this did not occur sufficiently to show up in the results, and requires further investigation.

4 Related work

There are very few reported attempts at help-desk response automation using purely corpus-based approaches, where the corpus is made up of request-response pairs. The retrieval system *eResponder* [Carmel *et al.*, 2000] retrieves a list of request-response pairs and presents a ranked list of responses to the user. This kind of document retrieval approach is also demonstrated by Bichel and Scheffer [2004], who also implement an approach similar to our *Doc-Pred*. Berger and Mittal [2000] present a summarization approach more akin to our *Sent-Pred*, but where a prediction model is learned directly from terms in requests and responses. The contribution of our work lies in the investigation of predictive approaches at different levels of granularity and the consideration of a hybrid prediction-retrieval approach.

5 Conclusion and Future Work

We have presented a predictive approach to the automation of help-desk responses, applied at two levels of granularity. Our methods 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. Our hybrid method was designed to overcome the loss of information resulting from abstracting response sentences. The use of sentence retrieval in combination with prediction was shown to be useful for better tailoring a response to a request. In future work, we intend to investigate a more focused retrieval approach that utilizes syntactic matching of sentences. For example, it may be beneficial to favour sentences that match on verbs. As another extension of our work we would like to improve the representation used for clustering, prediction and retrieval by using features that incorporate word-based similarity metrics [Pedersen *et al.*, 2004].

The results show that each of the prediction methods 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. A future avenue of research is thus to characterize situations where the different methods are applicable, in order to derive decision procedures that determine the best method automatically.

Our results suggest that the automatic evaluation method requires further consideration. Precision, and hence F-score, penalize good responses that are more informative than the model response. Our human judgements provide a more subjective indication of the quality of the generated response. However, a more extensive user study would provide an even more conclusive evaluation of the system, and could also be used to determine preferences regarding partial responses.

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 help-desk data.

References

[Berger and Mittal, 2000] A. Berger and V.O. Mittal. Query-relevant summarization using FAQs. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, pages 294–301, Hong Kong, 2000.

[Bickel and Scheffer, 2004] S. Bickel and T. Scheffer. Learning from message pairs for automatic email answering. In *Proc. of the European Conference on Machine Learning (ECML'04)*, Pisa, Italy, 2004.

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

[Filatova and Hatzivassiloglou, 2004] E. Filatova and V. Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *Proc. of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 397–403, Geneva, Switzerland, 2004.

[Lin and Hovy, 2003] C.Y. Lin and E.H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of the 2003 Language Technology Conference (HLT-NAACL'03)*, Edmonton, Canada, 2003.

[Marom and Zukerman, 2005] Y. Marom and I. Zukerman. Corpus-based generation of easy help-desk responses. Technical Report 2005/166, School of Computer Science and Software Engineering, Monash University, Clayton, Australia, 2005.

[Oliver, 1993] J.J. Oliver. Decision graphs – an extension of decision trees. In *Proc. of the Fourth International Workshop on Artificial Intelligence and Statistics*, pages 343–350, Fort Lauderdale, Florida, 1993.

[Pedersen *et al.*, 2004] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity – measuring the relatedness of concepts. In *Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI'04)*, pages 25–29, San Jose, California, 2004.

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

[Wallace and Boulton, 1968] C.S. Wallace and D.M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

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