21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

On the Feasibility of Answer Suggestion for Advice-seeking Community Questions about Government Services

Stephen Wan[a] **and Cécile Paris**[a]

[a]*CSIRO Data61*
Email: firstname.lastname@csiro.au

Abstract: Government departments are increasingly using social media and web resources to deliver information about services available to the citizenry. Part of this delivery includes online responses to community questions that are posted on social computing platforms such as Facebook, Twitter and forum sites. With the growth in the number of questions posted on such sites and the expectation that such questions are answered promptly, government customer service staff are faced with new challenges in terms of responding in a timely, accurate and complete manner.

In this work, we study a set of online engagements in which government customer service staff have answered community questions posted in public forums such as Twitter and Facebook pages. We examine the feasibility of automatically suggesting content for inclusion in the response, based on previous engagements. Ultimately, auto-suggestion of content may facilitate timely responses and help improve the consistency of answers.

We describe preliminary feasibility tests examining word overlap between question-answer pairs in our data set. Our results indicate that there is some degree of content overlap between question-answer engagements in our data set, suggesting that automatic suggestions for responses is possible.

Keywords: *Community question-answering, Gov 2.0, improving government services*

1 INTRODUCTION

To deliver public services effectively, government departments have explored different methods for providing accurate and timely information about the services. Often, the provision of such information requires one-to-one contact with a government representative who can help with questions about a service, especially when personal context is important. For example, questions about eligibility that depend on personal circumstances are often answered in call centres or online chat sessions. More recently, however, government departments are using social media platforms for this purpose, as questions and answers appropriate for a general audience can be viewed more widely when the information is indexed and presented via different web search engines.

An example is presented in Figure 1, in which a question is asked about eligibility criteria for a social service, which helps fund tertiary study. The question is originally answered by a member of the public. It is then answered by the organisation expert. Then, the original poster of the question has a follow up question which is again answered by the expert.[1]

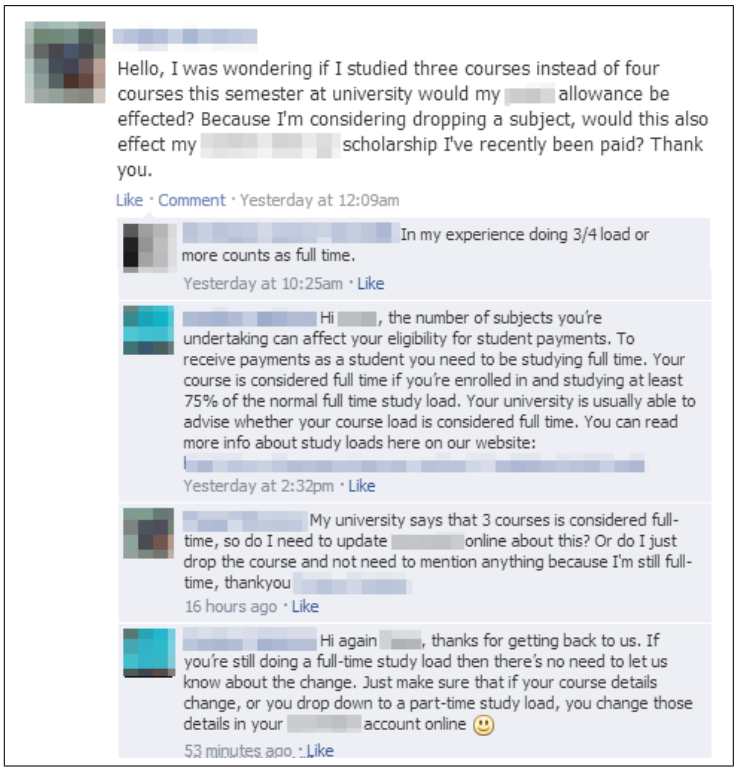

Figure 1. An example of a QA discussion from a managed Facebook page.

The kind of question-answer scenario we consider occurs on Facebook and Twitter. Specifically, moderated Facebook public pages and official Twitter accounts dedicated to government services are managed by government customer service staff as a focal point for receiving questions. As can be seen in Figure 1, questions posted on these platforms often include general descriptions of personal circumstances and, on these platforms, there is an opportunity for both the experts and the community to answer.

Social media can be described as "streaming data" and is typically presented using a "data feed" presentation mode in the user interface. Part of the issue of providing customer service via such interfaces is that old dialogues can drop off the feed, thus introducing a time-critical element to the task. This differs from community question-answering (cQA) platforms such as Yahoo!Answers[2] and StackExchange[3], where existing

[1] This example, based on a real discussion, has been shortened and anonymised.
[2] answers.yahoo.com
[3] stackexchange.com

question-answer pairs are always available via an index or through web search results, rather than as a data feed.

Our ultimate goal is to build a tool for suggesting content, particularly based on prior responses, to the government customer service staff for inclusion in a response to community questions. Given the sheer number of customers connected via social media, and the growing expectation that such questions be fielded online, such a tool has the potential to assist time-poor staff in customer engagements. Suggested content also has the added benefit of improving consistency of answers across the customer service team.

To examine the feasibility of such a tool, we investigate to what extent one can utilise previous question-answer pairs as such a resource. In this paper, we describe a preliminary investigation in which we examine the repetition of content between different question and answer pairs. Measuring text similarity using word overlap, we find that there is a degree of overlap that goes beyond simply recycling stock phrases.

2 Our user scenario

We are working with the Australian Government Department of Human Services (Human Services), which provides a variety of services for Australian citizens. Trained staff monitor social media to reply to advice-seeking questions. This activity happens on two Human Services-managed Facebook pages and an official account on Twitter. These engagements are archived, providing us with the question-answer data set used in this work.

In general, Human Services staff like to see community-generated answers as it fosters a supportive atmosphere in a forum. While community answers can be helpful, constant vetting by department staff is necessary as these answers can occasionally be out-of-date or incorrect. For such cases, a department representative would then engage with the online community to ensure that accurate information is provided.

We envisage an authoring tool to support this activity where, alongside the post detected as requiring a response, suggested content is presented. We investigate drawing this content from archived engagements. Ultimately, our aim is to build a tool that suggests relevant facts and key phrases extracted from the body of archived responses. In this work, we begin by seeing if there exists re-usable content, by examining word overlap between questions-answer pairs in this data set.

3 Related Work

We focus on questions about government services with some personal context. These are an example of advice-seeking requests, characterised by Mendes Rodrigues and Milic-Frayling (2009) as encompassing personal perspectives, concerning community and individual issues. Other categories of questions include requests for factual information and opinions (*ibid*), which have been more widely studied compared to our focus here. For example, Lui and Baldwin (2010) examines the social standing of the discussants on StackExchange, a forum for technical discussions on computing. Choi et al. (2012) argues that advice-seeking questions tend to occur in community-based platforms, such as Facebook pages, but not in expert-based platforms, such as StackExchange.

There are essentially two broad categories of related prior work on community question-answering, often from platforms such as Yahoo!Answers and StackExchange. The first is to locate an answer within the discussion, for some given question. This generally involves comparisons of different sentences in the discussion to the question, where similarity is variously defined. Catherine et al. (2012) found that lexical similarity between questions and answers can be low. To locate answers, Cong et al. (2008) introduced classification and graph-based methods, and Wang et al. (2010) applied a Deep Belief Network approach for this task. Hao et al. (2010) uses topical diversity as an alternative method to word overlap for scoring the similarity of questions.

The second category is about suggesting answers for questions. Our scenario differs from question-answering work in that most of this work has been applied to detecting answers for questions requiring a factual answer rather than advice-seeking questions (for an overview of question-answering for specific application domains, see Mollá and Vicedo (2007)). To some extent, our scenario has more similarity to the response automation methods, where incoming requests for help are classified to improve the routing of the request to the appropriate staff member (for an example of such work in the email domain, see Marom and Zukerman (2009)). In our case, the set of questions and topics is not fixed in advance, making text classification approaches difficult.

Finally, in suggesting content for answers, we recognise that there is a body of work on personalisation of generated text, including advice (for an example of personalised travel recommendations, see Colineau et al.

(2013)). However, we note that often such work is about tailoring content to suit an individual based on a general profile. In advice-seeking community question-answering, the personalisation is extremely fine-grained, and so text generation approaches based on user profiles, which are often coarse-grained in their handling of domain knowledge, may not be appropriate.

4 DATA

Our data set of engagements contains approximately 11,000 question-answer pairs in which questions are personal requests for advice, and answers have been provided by department experts. The data analysed occurred between the 25^{th} of March, 2013 and 1^{st} of May, 2015. For the analyses described here, data was first preprocessed by normalising text to lower casing and removing stop words. Questions and responses were then represented in vector space using TF-IDF weighting (using standard approaches as described in Salton and McGill (1983)).

Table 1 presents an overview of the statistics for our data set. Given the differences between the Facebook and Twitter platforms, we present the statistics separately. As expected, Facebook question-answer dialogues are longer than those of Twitter, given the latter's 140 character limit on the size of a post. We also note that there are many more pairs on Facebook than on Twitter.

Table 1. Descriptive statistics for the data.

Attribute	Facebook	Twitter
Q&A pairs	9134	2383
Unique words	8246	9615
Word count	205957	145759
Ave. Ques. length (words)	24	10
Ave. Ans.length (words)	44	12

5 A FEASIBILITY STUDY: LEXICAL OVERLAP

In this section, we ask a series of questions designed to characterise the extent to which there is lexical overlap that can help with content suggestion for responses.

5.1 Performance ceiling: question similarity

To begin with, we describe the similarity between questions in our data set. This might reveal whether it is possible to suggest content by identifying similar questions and then re-using their answers.

We use a sample of 1000 *test case* items taken from each platform. Each test case question is compared to the remaining questions in the data set, which we refer to as *the candidates*. Specifically, we ranked the candidate questions based on their cosine similarity with the test case question. To gauge content overlap, recall and precision of words was then calculated between questions at each rank level.

Figure 2 shows the overall similarity between test case and candidate questions. We observe that, at rank 1, recall and precision for lexical overlap are quite low, with only around a third of content-bearing words matching. This suggests questions are indeed expressing the variety of contexts idiosyncratic to the author, accounting for the reduced overlap in words.

5.2 Performance ceiling: response similarity

We repeat the process in Section 5.1, but instead we look at the similarity across answers. Since our data set provides a ground truth answer for a test case question, we want to understand how similar that response is to the other responses in the data set. This allows us to examine the upper bound of recommending words for responses.

The graph in Figure 3 shows the results for the lexical overlap of responses, using the same procedure as in Section 5.1. For either platform, we find that recall and precision of words (between responses) is approximately 50% at rank 1.

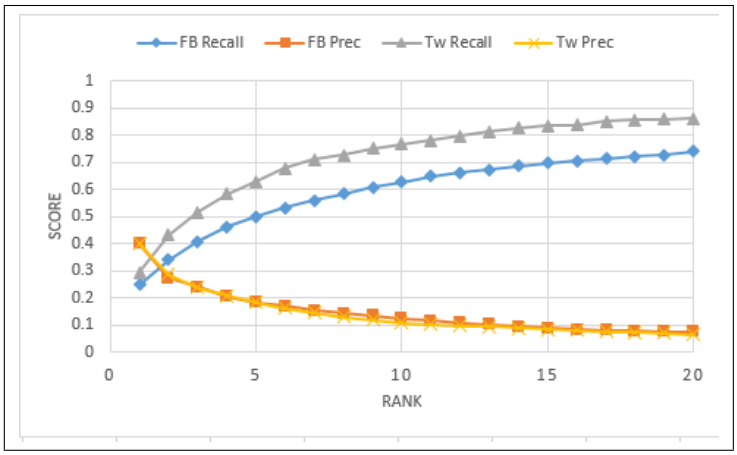

Figure 2. Upper bound on recall and precision of lexical items in the question.

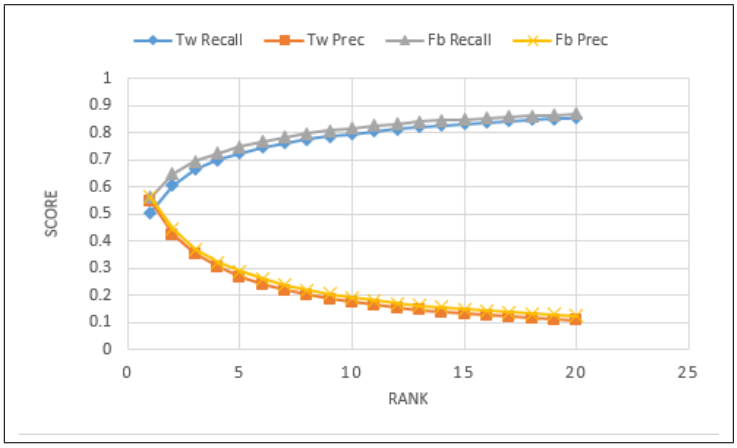

Figure 3. Upper bound on recall and precision of lexical items in the response.

We make two observations: (1) there is more similarity between responses than between questions; (2) although there are differences in conventions and lengths of posts on Twitter and Facebook, recall and precision performances are similar. We suspect that one normalising factor may be that answers are written by a small group of people employed at the department, which may account for this increased similarity.

This leads us to conclude that, at least in this context, there is scope for an authoring tool to help suggest content for the response and help with consistencies in content and style.

5.3 Accounting for boilerplate responses

We also examined whether the overlap in responses could be due to 'boilerplate' material replicated across different responses, looking at the recall and precision of content words using the top N words found in responses.

For Twitter, when using the top 12 words (the average length of a response), recall is 19%, whereas precision is 2%. For Facebook, the average length of a response is 44 words. For this number of words, we found a recall of 26% and a precision of 4%.

The recall and precision performance for this approximation to boilerplate content falls far below the ceiling of performance shown in section 5.2, from which we infer that boilerplate content does not account for the overlap. Rather, the overlap in content for responses does include content specific to the question.

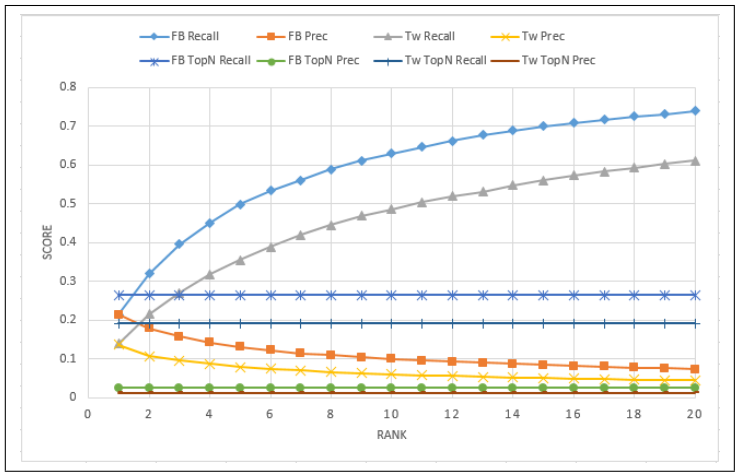

Figure 4. Recall and precision of suggested lexical items.

5.4 Answer suggestion by question similarity

Finally, we examine the performance of a simple answer suggestion mechanism by comparing the gold standard answer with the answer of the closest matching question. We first rank candidate questions to the test case question based on cosine similarity, from which we derive a ranking of candidate responses. Figure 4 shows the recall and precision of words for the different ranked responses. As a baseline, we also show the performance achieved by always returning the 20 most frequent words in the the data set of responses.

We find that there is a gap between the overall performance for this content suggestion method and the performance ceiling in Section 5.2, though it does beat the boilerplate baseline from rank 2 onwards. It is encouraging that even a simple answer suggestion mechanism may help, indicating that answer suggestion might be feasible with this data set, but it is also clear that there is room for improvement.

6 Conclusions and Future Work

Our experiments on the feasibility of an authoring tool for responses to personal advice-seeking questions indicate that the suggestion of response content is possible. That is, knowledge embedded in prior question-answer engagements appears to have some utility in formulating new responses. However, simple approaches to suggest content based on question similarity have limited applicability, falling short of the upper bound in performance. In future work, we will investigate Named Entity Recognition (NER) approaches that will help identify longer spans of text to suggest as content for an answer. We also intend to explore different approaches to bridge the lexical gap between the question to the response content as outlined by Catherine et al. (2012). Recent approaches using word embeddings may provide a richer projection in which to determine the similarity of questions and responses. Finally, we note that the department's activities for answering Facebook and Twitter content are currently performed independently of each other. In future work, we hope to explore how answers on one platform may help with questions posted on the other.

Acknowledgments

This work builds on work partially funded under the Human Services Delivery Research Alliance between the CSIRO and the Australian Government Department of Human Services. We thank Human Services for letting us use their data for this research. We also thank the reviewers of the original paper for their constructive comments.

References

Catherine, R., A. Singh, R. Gangadharaiah, D. Raghu, and K. Visweswariah (2012, December). Does similarity matter? the case of answer extraction from technical discussion forums. In *Proc. of COLING 2012: Posters*, Mumbai, India, pp. 175–184. The COLING 2012 Organizing Committee.

Choi, E., V. Kitzie, and C. Shah (2012). Developing a typology of online Q&A models and recommending the

right model for each question type. *ASIST 49*(1), 1–4.

Colineau, N., C. Paris, and K. Vander Linden (2013). Automatically producing tailored web materials for public administration. *New Review of HyperMedia and MultiMedia 9(2)*, 158–181.

Cong, G., L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun (2008). Finding question-answer pairs from online forums. In *Proc.of SIGIR'08*, SIGIR '08, New York, NY, USA, pp. 467–474. ACM.

Hao, T., W. Liu, and E. Agichtein (2010, June). Towards automatic question answering over social media by learning question equivalence patterns. In *Proc.of the NAACL HLT 2010 Wkshp on Computational Linguistics in a World of Social Media*, Los Angeles, California, USA, pp. 9–10. Association for Computational Linguistics.

Lui, M. and T. Baldwin (2010, December). Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet. In *Proc. ALTA 2010*, Melbourne, Australia, pp. 49–57.

Marom, Y. and I. Zukerman (2009). An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Computational Linguistics 35*(4), 597–635.

Mendes Rodrigues, E. and N. Milic-Frayling (2009). Socializing or knowledge sharing?: Characterizing social intent in community question answering. In *Proc. of CIKM 2009*, CIKM '09, New York, NY, USA, pp. 1127–1136. ACM.

Mollá, D. and J. L. Vicedo (2007, March). Question answering in restricted domains: An overview. *Computational Linguistics 33*(1), 41–61.

Salton, G. and M. J. McGill (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Wang, B., X. Wang, C. Sun, B. Liu, and L. Sun (2010, July). Modeling semantic relevance for question-answer pairs in web social communities. In *Proc. of ACL 2010*, Uppsala, Sweden, pp. 1230–1238. ACL.