Alfalahi, Alyaa, Gunnar Eriksson, and Eriks Sneiders. "Shadow Answers as an Intermediary in Email Answer Retrieval." *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer International Publishing, 2015. 209-214. http://link.springer.com/chapter/10.1007/978-3-319-24027-5_18

Shadow Answers as an Intermediary in Email Answer Retrieval

Alyaa Alfalahi[1], Gunnar Eriksson[1], Eriks Sneiders[1]

[1] Stockholm University, Department of Computer and Systems Sciences
Postbox 7003, SE-164 07, Kista, Sweden
{alyalfa, gerik, eriks}@dsv.su.se

Abstract. A set of standard answers facilitates answering emails at customer care centers. Matching the text of user emails to the standard answers may not be productive because they do not necessarily have the same wording. Therefore we examine archived email-answer pairs and establish query-answer term co-occurrences. When a new user email arrives, we replace query words with most co-occurring answer words and obtain a "shadow answer", which is a new query to retrieve standard answers. As a measure of term co-occurrence strength we test raw term co-occurrences and Pointwise Mutual Information.

Keywords: Email answering, statistical word associations, shadow answer.

1 Introduction

Agents at customer care centers traditionally use standard answers (a.k.a. answer templates) to answer customer emails. Various methods for obtaining email answers may help with this task. Matching manually crafted text patterns yields the highest accuracy of answer retrieval [1], but it is a labor intensive approach. Machine learning is popular (e.g. [2-3]), but it works best with a few and broad text categories. Answer generation (e.g. [4]) is an interesting research problem, but not likely to reach commercial use in the nearest future.

Our contacts with customer care centers in Sweden show that they prefer technology support that requires minimum maintenance, and this minimum does not depend on rare professional competence. For email answering that means a focus on statistical text similarity calculation rather than building a knowledge base (e.g. [5]).

Our task at hand is retrieval of standard answers when a new customer email arrives. The difficulty of the task is different wordings: a standard answer is not a document similar to the query, it is a document that answers the query. We cannot rely on term similarity. There exist, however, statistical word associations: certain words in similar queries co-occur with certain words in their answers. This may be a machine learning task for Support Vector Machine (SVM). Alternatively, we can measure the strength of these associations and use them in order to replace words in a user email with the associated words from the answers. Thus, the user email is translated into a shadow answer, i.e., a user query made of anticipated answer words, which becomes a

new search query in the database of standard answers. The question is – how can we measure the word associations between user emails and their answers? We compare two measures – raw term co-occurrence and Pointwise Mutual Information.

Further in this paper, Section 2 presents our answer retrieval method. Sections 3 and 4 introduce the experiment data and process. Section 5 shows the results, and Section 6 concludes the paper.

2 Shadow Answer

Because we cannot use the original user email as a search query among standard answers, we translate the user email into a shadow answer that contains terms expected in the answer, and use the shadow answer as a search query among standard answers. The idea of a shadow answer comes from Lamontagne et al. [6] who explored co-occurrences between words in archived problem descriptions and their solutions. Our messages and their answers are two parallel corpora; parallel corpora are traditionally used in machine translation to train the system to establish relationships between similar words in two languages. We have a similar task; our "two languages" are the wording of user emails and the wording of their answers.

The architecture of our answer retrieval process is shown in Fig. 1.

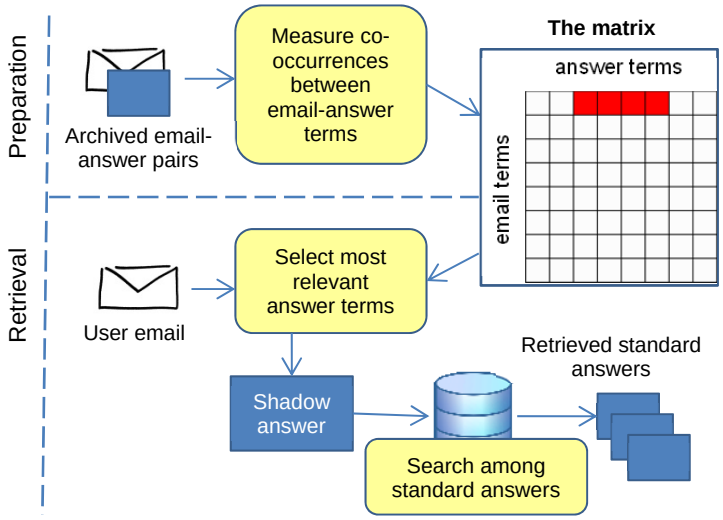

Fig.1. Answer retrieval by translating user query terms into answer terms.

During the preparation phase, we measure term co-occurrences in archived emails and their answers, and fill the numeric co-occurrence values into the matrix. Every term in the email corpus has a corresponding row in the matrix; every term in the answer corpus has a corresponding column in the matrix. The numeric values in the matrix

show the strength of co-occurrence of two terms in the email and its answer respectively.

During the answer retrieval phase:

1. The system takes each term in the user email and consults the matrix for one or several most co-occurring answer terms, and puts these answer terms into the shadow answer, which is a bag of words. If an email term has no corresponding answer term, it is ignored. The shadow answer is an equivalent of the user email re-written in answer terms.
2. We use the shadow answer as a search query for a standard text-retrieval system to get a ranked list of standard answers.
3. Because the shadow answer contains terms expected in the answer of the given user email, we hope that the retrieved answers are relevant.

Our research question is how we can fill and use the matrix in Fig. 1. In this paper, we explore two measures of term co-occurrence. First one is raw co-occurrence, i.e., the number of email-answer pairs where one term occurs in the email and the other term occurs in the answer. Second one is Pointwise Mutual Information (PMI).

PMI is a simple measure of co-occurrence strength between two items. It works by relating the probabilities of the individual occurrence of the items to the probability of both items occurring together. In this paper, the probabilities of query and answer term are based on their occurrence in all questions and answers, respectively. The joint probability of the co-occurrence of a pair of a question term and an answer term is based on their occurrence in the same question-answer pair. For more information on this measure, see e.g. Yang and Pedersen [7].

Our goal is to find out whether PMI is better than the raw term co-occurrence for generating shadow answers.

3 Experiment Data

Our data is 1431 email-answer pairs from the Swedish Pension Authority (Pensionsmyndigheten in Swedish). Because we had a text retrieval task, not a traditional machine learning task, we did not divide our email collection into training and test data. We used all 1431 email-answer pairs to fill the email-answer term co-occurrence matrix.

During the answer retrieval test, we used all 1431 emails as user emails, and all 1431 answers as simulated “standard answers”. We increased the number of test answer texts by adding some FAQ answers from the Pension Authority’s homepage.

4 Experiment Process

Measuring co-occurrences between email-answer terms. Two parallel sets of experiments were conducted. One set of experiments filled the email-answer term co-

occurrence matrix with raw term co-occurrence values; the other set of experiments had the matrix filled with term PMI scores. The texts were not stemmed or lemmatized. Separate sub-experiments were conducted with and without removal of stop-words from user emails and their answers.

Selecting most relevant answer terms and generating a shadow answer was conducted roughly the same way when using term PMI or raw term co-occurrences. The system took each email term, consulted the matrix, selected most co-occurring answer terms, and put those answer terms into the shadow answer.

Search among standard answers. The shadow answer becomes a query for Lucy, our text-retrieval system. Lucy (http://lucy.apache.org/) is an open source information retrieval system with a standard tf-idf-based ranking. In our experiments, document indexing was performed with Swedish stemming, but without any other modifications such as stop word filtering.

Retrieval performance measurements. At the moment of conducting the experiments, the only proof of email-answer relevance was the fact that both the email and the answer originally were in the same pair. We do not formally know whether the answer in a different email-answer pair is relevant to the given email or not, although in reality there are many similar answers. We measured the retrieval performance as follows:

- Lucy retrieved a ranked list of answers.
- In the list of answers, we looked for the original answer of the submitted user email; i.e., they both originally were in the same email-answer pair.
- We note the rank, i.e., the position in the list, of the original answer.
- The average rank of original answers across all 1431 submitted emails describes the potential of the retrieval method.

Baseline method. Our baseline method was submitting the email message directly to the text retrieval system without the matrix and the shadow answer. The baseline method searched for answers similar to the text of the user email.

5 Experiment Results

Table 1 shows the answer retrieval results when we filled the email-answer term co-occurrence matrix with raw term co-occurrences. The last row shows the results of the baseline method – no matrix at all.

The first four rows in the table stand for sub-experiments: for each term in the submitted email we selected top *n* most often co-occurring answer terms to put into the shadow answer.

The second and third columns stand for another kind of sub-experiment: when the matrix was filled, stop-words were left in the text or removed from the text.

The cells of the table show the average rank of the original answer across all the submitted emails.

Table 1. Answer ranks, the matrix filled with raw term co-occurrence

Top *n* co-occurring	Avg. rank with stop-words	Avg. rank without stop-words
Top 1	431	202
Top 5	327	239
Top 20	256	304
Top 30	293	320
Baseline	463	184

The biggest surprise is the low rank of the original answers in the list of retrieved answers – the highest average is 184. Because we use a mixture of techniques, we cannot blame any single technique for that. The next biggest surprise is the baseline method, which is the best performing method if stop-words are removed from the texts. If we do use the shadow answer, it is better to remove stop-words and select fewer top co-occurring answer terms.

Table 2 shows the answer retrieval results when we filled the email-answer term co-occurrence matrix with term PMI scores. We extended our PMI experiments by using not only unigrams but also bigrams, terms made of two consecutive words. The email-answer term co-occurrence matrix was filled once by PMI scores between unigrams, bigrams, as well as between unigrams and bigrams.

During the retrieval, we selected only top 1 co-occurring answer terms to be placed into the shadow answer, which corresponds to the first row of Table 1. Furthermore, we experimented with selecting only unigrams, only bigrams, or both, in the user email, and putting only unigrams, only bigrams, or both, into the shadow answer. In Table 2, "$U_e \rightarrow U_{sa}$" stands for the experiment where unigrams were selected in the user email, and unigrams were placed into the shadow answer, as in the experiments in Table 1. "$U+B_e \rightarrow U+B_{sa}$" means that both unigrams and bigrams were selected in the user email, as well as both placed into the shadow answer; the co-occurrences between unigrams, bigrams, and between unigrams and bigrams were considered.

Not surprisingly, the best gain was from using longer sequences, i.e. bigrams: the best average rank of the original answer was obtained by selecting only bigrams from user emails and putting only bigrams into the shadow answers. On the other hand, mixing unigrams with bigrams performed worst, as the last row in Table 2 shows.

Table 2. Answer ranks, the matrix filled with term PMI scores

Selection of uni/bi-grams	Avg. rank	Selection of uni/bi-grams	Avg. rank
$U_e \rightarrow U_{sa}$	66	$B_e \rightarrow U+B_{sa}$	49
$B_e \rightarrow B_{sa}$	28	$U_e \rightarrow U+B_{sa}$	68
$U_e \rightarrow B_{sa}$	38	$U+B_e \rightarrow U_{sa}$	78
$U+B_e \rightarrow B_{sa}$	47	$U+B_e \rightarrow U+B_{sa}$	81
$B_e \rightarrow U_{sa}$	48		

6 Conclusions

The concept of a shadow answer is not new, yet barely used in answer retrieval. We believe this concept has a potential together with a good measurement of term co-occurrences. In our experiments, term PMI outperformed raw term co-occurrence. In experiment settings where only unigrams were used, PMI yielded 66 as the average rank of the original answer, while raw term co-occurrence yielded 202. Having the original user email as the search query (i.e., as the shadow answer) in the set of answers yielded the average rank 184. The best average rank – 28 – was achieved with PMI and bigrams.

We had an unusual method for measuring the performance of answer retrieval – the rank (i.e., position) of the original answer of the user email in the list of retrieved answers. We chose this method because we did not have expert-labeled documents as it is common in text retrieval evaluation. The average rank turned out to be much lower than we expected, although we saw relevant documents on the top of the answer list. For practical use, it appears that shadow answer alone may not be sufficient. Our ongoing research suggests that it can be used in a combination of retrieval methods that generates a merged result list.

We are in the process of labelling answers, which would allow us improving future relevance judgements.

References

1. Sneiders, E.: Automated Email Answering by Text Pattern Matching. In: H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (eds.): Proc. 7th International Conference on Natural Language Processing (IceTAL), August 16-18, Reykjavik, Iceland, LNAI 6233, pp. 381-392. Springer, Heidelberg (2010)
2. Lapalme, G., Kosseim L.: Mercure: Towards an automatic e-mail follow-up system. In: IEEE Computational Intelligence Bulletin 2, no. 1, pp. 14-18. IEEE (2003)
3. Itakura, K., Kenmotsu, M., Oka, H., Akiyoshi, M.: An identification method of inquiry e-mails to the matching FAQ for automatic question answering. In: Distributed Computing and Artificial Intelligence, pp. 213-219. Springer Berlin Heidelberg (2010)
4. Marom, Y., Zukerman, I.: Towards a framework for collating help-desk responses from multiple documents. In: Proceedings of the IJCAI05 Workshop on Knowledge and Reasoning for Answering Questions, pp. 32-39 (2005)
5. Malik, R., Subramaniam, L.V., Kaushik, S.: Automatically Selecting Answer Templates to Respond to Customer Emails. In: IJCAI, vol. 7, pp. 1659-1664 (2007)
6. Lamontagne, L., Langlais, P., Lapalme, G.: Using Statistical Word Associations for the Retrieval of Strongly-Textual Cases. In: FLAIRS Conference, pp. 124-128 (2003)
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 412-420 (1997)