Clustering E-Mails for the Swedish Social Insurance Agency – What Part of the E-Mail Thread Gives the Best Quality?

Hercules Dalianis[1], Magnus Rosell[1,2], and Eriks Sneiders[1]

[1] Department of Computer and Systems Science,
(DSV) Stockholm University
Forum 100, 164 40 Kista, Sweden
[2] KTH CSC, 100 44 Stockholm, Sweden
hercules@dsv.su.se,rosell@csc.kth.se,eriks@dsv.su.se

Abstract. We need to analyse a large number of e-mails sent by the citizens to the customer services department of a governmental organisation based in Sweden. To carry out this analysis we clustered a large number of e-mails with the aim of automatic e-mail answering. One issue that came up was whether we should use the whole e-mail including the thread or just the original query for the clustering. In this paper we describe this investigation. Our results show that only the query and the answering part should be used, but not necessarily the whole e-mail thread. The results clearly show that the original question contains more useful information than only the answer, although a combination is even better. Using the full e-mail thread does not downgrade the result.

Keywords: E-government, query answering, e-mail threads, Swedish, clustering.

1 Introduction

In Sweden the public authorities have been in the lead to implement E-government. This includes communication with the citizens through various electronic channels. One such channel is to put important information on their web sites. Citizens often do not find the information they are seeking, however, and initiate communication in one of several ways, such as telephone calls, e-mails, chat lines, etc.

The Swedish Social Insurance Agency[1] (SSIA) receives more than 10 000 e-mails from citizens each week. These are answered manually by handling officers. Many of the e-mails from the public are very similar. Therefore a lot would be gained if these re-occurring questions could be answered automatically or semi-automatically. To accomplish this, first the common questions must be identified.

The e-mails are either sent directly to an available address or via a web form on the agency's web site. When a citizen uses the web form he/she also has to assign

[1] www.forsakringskassan.se

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 115–120, 2010.

a category to it, such as parental benefit (föräldrapenning), housing allowance (bostadsbidrag), superannuation (pension), sickness benefit (sjukpenning), etc.

Although these broad categories help to assign the e-mails to the right handling officer they do not help to identify the common questions. To find groups of common questions we apply text clustering.

Our ultimate goal is to help the handling officers to use clustering as a tool to facilitate more efficient and up-to-date answers. Clustering could be used to get an overview of the trends in the questions and to identify common questions that could be answered using a standard answer. In the long run such questions could also be answered automatically, [1].

In the present work we investigate clustering of the e-mails without involving the handling officers. We study the effect of using different parts of the e-mail threads in order to achieve the best clustering quality.

2 Previous Research

An e-mail consists of a header (including sender and receiver addresses, subject matter, etc) and body text. The body text may also be divided into several zones of different kinds of content, such as sender zones *(author, greeting, signoff)*, quoted conversation zones *(reply, forward)*, and boilerplate zones *(signatures, advertising, disclaimer, attachment)* [2].

Previous work on clustering of e-mails has discussed the inclusion of different parts of the e-mails, but has not tried different parts of the body. In [3] using a combination of the header and body gives better results than using only the body. In [4] the authors let the user weight the importance of the parts *(to, cc, from, subject, date, body)*.

Whereas previous research was aimed at personal inboxes, we study e-mails sent to a whole organisation.

3 Text Sets and Preprocessing

We received about 9 000 e-mails from the SSIA. Around 4 000 of these were either sent directly (without the use of the web form) or assigned a miscellaneous category “other questions” (övriga frågor) in the web form. As we could not use these for our evaluation we removed them, producing a set of almost 5 000 e-mails that were categorised.

All e-mails were also de-identified because of their sensitive nature. The de-identification of the e-mails was carried out by SSIA before the e-mails were handed over to our research group. The de-identification program was developed and evaluated by our research group. For first names we obtained an F-score of 0.82 and a recall of 0.73 respectively and for last names an F-score of 0.85 and a recall of 0.77 respectively. For social security numbers, phone numbers, e-mail addresses, web addresses, street addresses and postal codes we obtained an F-score of 0.93 and a recall of 0.92.

3.1 Extracting Parts of the E-Mail Thread

The e-mails we obtained were actually complete e-mail threads as they had developed up until the moment they were extracted at the SSIA. The number of items in a thread varied from one to 40 although 96.2 percent of all threads where up to four components long.

The principle of separating thread components was empirically obtained by working on a large number of e-mails. The system iteratively cuts off the top message. It first looks for several successive lines that start with ">".

If these are found, then everything above these lines is the top message. Otherwise the system looks for a typical message separator line, such as "abc@doc.com wrote:", "Original message:", etc in several languages (Swedish, English, Norwegian) with a certain level of wording freedom. If this does not help, it looks for an array of lines that start with "From:", "To:", "Date:", "Subject:" in different languages. This method is based on heuristics but works comparably well.

For our clustering experiments we created four sets of texts: *Question* – a set containing only the first question in each thread, *Answer* – the first answer to each question, *Question and Answer* – the first question and the first answer, and *Thread* – the whole e-mail thread.

3.2 Lemmatisation and Filtering

Using a few simple rules we removed the e-mail headers and characters indicating quotation/citation of previous messages in the thread. We were not allowed to use the headers due to the sensitive nature of these e-mails.

The results for each of the different sets were tokenised and lemmatised using the Swedish grammar checking program Granska [5]. The resulting texts still contained a lot of non-word character sequences, coming from signatures, advertisements, disclaimers, etc. To try to remove them we have used several simple methods. We removed words shorter than three characters and longer than 20, since this only removes a few interesting words and captures some of the non-words. Further, we removed all words only appearing in only one e-mail, (see appendix in [6]), since they did not contribute to the similarity between e-mails. We also used a common stoplist of Swedish words.

3.3 Statistics for the Preprocessed Text Sets

Table 1 gives some statistics for the extracted and preprocessed text sets: the number of texts and lemmas, as well as the average number of different lemmas per text and the average number of texts in which each lemma occurs.

4 Clustering

For each text set we constructed an ordinary term-document-matrix with tf*idf-weights. We defined similarity between texts as the cosine measure.

We have used the K-Means algorithm, (see for instance [7]), as it is simple, fast, and therefore suitable for interactive exploration. In the end we want the handling officers to use clustering as a tool to obtain an overview of the trends in the questions and to indentify common questions, this in an interactive manner as described in [8].

5 Evaluation

Since internal clustering quality measures are based on the representation we can not use them to compare results based on different representations, i.e. our text sets. External quality measures compare the clustering with a categorisation. We have the categorisation made by the citizens. It may not be ideal, but at least it groups questions with similar content. We want clusterings to compare well with this categorisation, although we do not expect them to be very similar. We prefer a clustering to be more similar rather than less similar, however.

There are many external quality measures. We prefer information theory based measures as these take the whole distribution of texts over categories and clusters into account. For this reason we use the Normalised Mutual Information (NMI) between the clustering and the categorisation, see [9].

Table 1. Clustering results for four different text sets (based on the original question only, the first answer only, both first question and answer, and the full e-mail thread). The first four measures describe the text sets after preprocessing. The last measure is the average clustering result in NMI (Normalised Mutual Information) of 20 K-Means clusterings to nine clusters compared with the categorisation. Standard deviations are shown in parenthesis.

	Text Set			
Measure	Question	Answer	Question and Answer	Thread
Number of Texts	4 652	4 681	4 839	4 841
Number of Lemmas	2 929	2 055	3 956	4 398
Lemmas/Text	12.2	9.2	19.5	23.2
Texts/Lemma	19.3	21.0	23.9	25.5
NMI	0.28 (0.03)	0.14 (0.02)	0.40 (0.03)	0.38 (0.04)

6 Experiments and Discussion

In Table 1 we report average results in NMI for nine to 20 clusterings of the different text sets, with the standard deviation shown in parenthesis. In order for two results to be considered different they, as a rule of thumb, they need not overlap with their standard deviations.

We choose nine clusters as the categorisation has nine categories. The tendencies we describe are similar for other numbers of clusters.

The result clearly shows that the textual information in the question (*Question*) is better than what is in the answer (*Answer*). The result gets even better, however, if we also include the answer (*Question and Answer*). The result for the entire e-mail thread (*Thread*) is the same as for *Question and Answer.* As shown in Table 1 the Normalised Mutual Information (NMI) for the query and answering part is 0.12 units higher than for only the query alone.

It is not surprising that the result is better for the set of questions than for the set of answers as the categories are chosen by the citizens who also formulated the questions. The answers are often shorter than the questions (see the statistics in Table 1), use a more formal language, and do not necessarily include the same terms as their corresponding question. This makes the answers harder to group. Combined with the question, however, the answer does give more information (than using only the question) for the clustering algorithm to work with as similar questions tend to be answered in similar ways.

By the same reasoning, the result when using the entire thread is used should be even better. The questions that require more responses (follow-up questions with answers), however, are probably more complicated and therefore harder to group. The categorisation of the first question might not even be suitable for the entire thread as it may well include new questions regarding other matters.

As the full thread (*Thread*) contains most information and it performs equally well with the questions and answers (*Question and Answer*) we will use it in our further work.

7 Conclusions and Future Work

We have compared clusterings of e-mails sent to the SSIA based on different parts of the e-mail thread/texts. The results clearly show that the original question contains more useful information than only the answer, although a combination is even better. Using the full e-mail thread does not downgrade the result.

We plan to involve the handling officers in our next investigation. We will let them explore clusterings of the e-mails and interview them to learn whether an approach like this is actually useful and if it can provide insights, help to find common questions and formulate standard answers.

Acknowledgements. We would like to thank Anne-Lie Karlsson at Försäkringskassan, SSIA, for her devoted support of the IMAIL research group. We would also like to thank VINNOVA (The Swedish Governmental Agency for Innovation Systems) for the funding of the IMAIL-project.

References

1. Knutsson, O., Pargman, T., Dalianis, H., Rosell, M., Sneiders, E.: Increasing the efficiency and quality of e-mail communication in e-Governmnent using language technology. In: Proc. of IFIP e-Government Conference 2010 (EGOV 2010), Lausanne, Switzerland, August 29-September 2 (2010) (to be published)
2. Lampert, A., Dale, R., Paris, C.: Segmenting email message text into zones. In: Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009 (2009)

3. Huang, Y., Govindaraju, D., Mitchell, T.M., de Carvalho, V.R., Cohen, W.W.: Inferring ongoing activities of workstation users by clustering email. In: CEAS – Conference on Email and Anti-Spam (2004)
4. Schuff, D., Turetken, O., D'Arcy, J.: A multi-attribute, multi-weight clustering approach to managing “e-mail overload". Decision Support Systems 42, 1350–1365 (2006)
5. Domeij, R., Knutsson, O., Carlberger, J., Kann, V.: Granska – an efficient hybrid system for Swedish grammar checking. In: Proc. 12th Nordic Conf. on Comp. Ling. – NODALIDA 1999 (1999)
6. Rosell, M.: Text Clustering Exploration – Swedish Text Representation and Clustering Results Unraveled. PhD thesis, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden (2009)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
8. Cutting, D.R., Pedersen, J.O., Karger, D., Tukey, J.W.: Scatter/Gather: A cluster-based approach to browsing large document collections. In: Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (1992)
9. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617 (2003)