Automatic Summarization of Domain-specific Forum Threads: Collecting Reference Data

Suzan Verberne
Radboud University
s.verberne@let.ru.nl

Antal van den Bosch
Radboud University
a.vandenbosch@let.ru.nl

Sander Wubben
Tilburg University
s.wubben@uvt.nl

Emiel Krahmer
Tilburg University
E.J.Krahmer@uvt.nl

ABSTRACT

We create and analyze two sets of reference summaries for discussion threads on a patient support forum: expert summaries and crowdsourced, non-expert summaries. Ideally, reference summaries for discussion forum threads are created by expert members of the forum community. When there are few or no expert members available, crowdsourcing the reference summaries is an alternative. In this paper we investigate whether domain-specific forum data requires the hiring of domain experts for creating reference summaries. We analyze the inter-rater agreement for both datasets and we train summarization models using the two types of reference summaries. The inter-rater agreement in crowdsourced reference summaries is low, close to random, while domain experts achieve a considerably higher, fair, agreement. The trained models however are similar to each other. We conclude that it is possible to train an extractive summarization model on crowdsourced data that is similar to an expert model, even if the inter-rater agreement for the crowdsourced data is low.

Keywords

summarization, discussion forums, reference data

1. INTRODUCTION

Discussion forums on the web come in many flavors, each forum covering its own topic and having its own community. The user-generated content on web forums is a valuable source for information. However, in discussion forums where opinions and experiences are shared, it can be difficult to pinpoint the relevant information in a long thread, especially when the forum is accessed on a mobile device.

The Dutch media company Sanoma wants to serve its mobile users better by showing summaries of long discussion threads. The approach we take in this paper is *extractive* summarization [5]: extracting salient units of text from a document and then concatenating them to form a shorter version of the document. In previous work on extractive summarization for discussion threads it was assumed that *posts* are more suitable summarization units than sentences, because selecting sentences from posts would lead to loss of context [1]. However, posts are sometimes long and not all information contained in a post is equally relevant. In this paper we investigate the feasibility of thread summarization at the sentence level: selecting the most relevance sentences from a thread while hiding the less relevant sentences.

For the development and evaluation of automatic summarizers, human reference summaries are required [11]. However, summarization is an inherently subjective task: human summarizers tend to disagree on the information that should be included in the summary. When multiple raters are involved in creating the reference summaries, it is insightful to measure the inter-rater agreement between them. Agreement for extractive summarization – the selection of salient text units from a document – is generally calculated in terms of Cohen's κ [13]. For summaries of newswire texts κ scores between 0.20 and 0.50 have been reported [10], but for the summarization of conversations the agreement tends to be lower: between 0.10 and 0.35 [9, 7, 12].

The case that we study in this paper is that of web forums of patient support groups. Online support groups play an important role in providing patients with informational and emotional support from their peers [4]. The information shared in patient support groups is domain-specific, covering the disease, medications, side effects and coping strategies. Ideally, reference summaries for discussion forum threads are created by active members of the forum community, who are experience experts in the domain of the forum and familiar with the commonalities of the forum community. When there are few or no expert members available, crowdsourcing the reference summaries is an alternative. In this paper we investigate whether domain-specific forum data requires the hiring of domain experts for creating reference summaries, or that crowdsourced reference summaries suffice. We collect two sets of reference summaries for a set of discussion threads from the Facebook group of a patient support community: expert summaries and crowdsourced summaries. We analyze the inter-rater agreement for both datasets, and compare the models trained on both datasets. We address the following research questions:

RQ1 What is the inter-rater agreement for sentence selection from domain-specific discussion forum threads;

CHIIR '17, March 07-11, 2017, Oslo, Norway

DOI: http://dx.doi.org/10.1145/3020165.3022127

do domain experts achieve a higher agreement than crowdsourced raters on the task?

RQ2 What are the differences between sentence selection models trained on expert data and crowdsourced datas?

We release the labeled data to be reused by other researchers.[1]

2. RELATED WORK

Over the last decade, some research has been directed at the summarization of forum threads [14, 8, 3]. The most similar to the current work in the context of discussion forum summarization is the work by Bhatia et al. [1], which takes a feature-based approach in selecting the most relevant posts from a thread, arguing that sentence selection would lead to loss of coherence.

The quality of crowdsourced annotations compared to expert annotations has been investigated before in the context of other supervised learning tasks. For the classification of sentiment in political blog snippets it was found that having one individual non-expert annotator leads to a model with poor quality, but using multiple noisy annotations from non-experts can lead to a useful model [6]. For relevance judgments in domain-specific search, crowdsourced judgments give the same results as expert judgments in a global evaluation (ranking of evaluated search engines), but fail in more specific evaluation tasks (distinguish different levels of highly accurate search results) [2].

This paper is the first to analyze the difference between expert annotations and crowd-sourced annotations for the summarization of discussion forum threads, and the first to study the potential of sentence-level summarization of discussion forum threads.

3. METHODOLOGY

3.1 Data

We use data from the Facebook group GIST support international.[2] GIST is a rare form of cancer. We received an export of 3,071 discussion threads from the Facebook group. Threads with more than 26 posts were cut off at 26 by the provider of the data. The average number of posts in a thread after cut-off is 9.5 (median 7). 40% of the threads have more than 10 posts. For our experiment we created a sample of 100 randomly selected threads that have at least 10 posts. We automatically split each post in sentences, using regular expressions and a set of abbreviations in Python. The average number of sentences per post is 2.7 (minimum 1, maximum 18).

3.2 Creating reference summaries

For creating reference summaries, we set up an online thread summarization interface. In this interface the left column of the screen shows the complete thread. When hovering over the text in the posts, separate sentences get highlighted and are clickable. The right column shows an empty table with a placeholder cell for each post and within each post a placeholder `[...]` for each sentence in the post. The subjects were given the following instructions: *"Please select the pieces of text that you think are the most important for the thread. You can either select a post as a whole (by clicking 'select the complete post') or select separate sentence(s) from a post. You can determine the number of selected posts yourself but try to be concise so that the resulting summary does not contain too much redundant information. The selected text together should form an informative summary of the thread."* The subjects also had the possibility to remove sentences or posts from the selection by clicking the selected items.

With the annotation interface, we collected two sets of reference summaries:

- **Crowdsourced summaries.** Through the research participation system of our university, we recruited subjects to create reference summaries. The users provided some basic information in the login screen, such as their gender and age. They were then presented with one example thread to get used to the interface. After that, they were presented with a randomly selected thread from our sample. The subjects decided themselves how many threads they wanted to summarize. They were paid a gift certificate. Each thread was shown to 5 different subjects.
- **Expert summaries.** Through the GIST patient community, we recruited 2 experts to create reference summaries. 50 of the 100 threads in our sample were summarized by the experts, using the same instructions and annotation interface as for the crowdsourced summaries.

Table 1: Sentence features used as independent variables in the regression analysis. Some of the features describe the post in which the sentence occurs.

Name	Description
pos_post	position of post in the thread
pos_sent	position of sentence in the post
cossim_thread	cosine similarity sentence–thread[1]
cossim_openingpost	cosine similarity sentence–opening post[1]
wordcount	word count
ttr	type-token ratio
rel_punctcount	relative punctuation count
avg_wordlength	average word length (# of chars)
author	proportion of posts by author of post

1. tf-idf weighted term vectors

3.3 Inter-rater agreement

For each thread we computed the agreement between each pair of raters in terms of Cohen's κ.We report the mean κ scores over all threads and all rater pairs.

3.4 Training a model for sentence selection

In order to answer RQ2 we investigated the relationship between sentence features and the selection of sentences using a linear regression analysis. We argue that the number of raters that selected a sentence (the number of 'votes') is an indicator of its relevance: a sentence that is selected by all 5 raters can be expected to be more relevant than a sentence that is selected by only one or two raters. Therefore, we used the number of votes for a sentence as dependent variable in the regression analysis. The sentence features that we used as independent variables are listed in Table 1. We standardized feature values by converting them to their z-value.

[1]The data can be downloaded from http://discosumo.ruhosting.nl/

[2]https://www.facebook.com/groups/gistsupport/

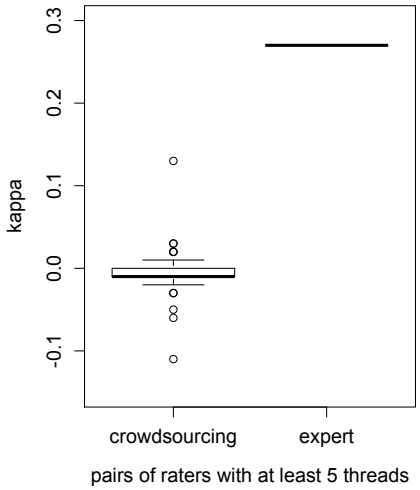

Figure 1: Dispersion of κ scores for the rater pairs with at least 5 common threads (κ scores are averages over the κ scores for individual threads).

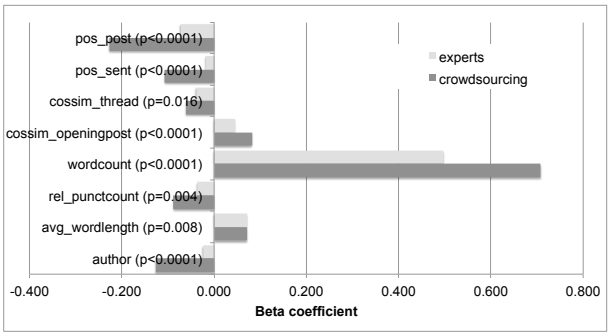

Figure 2: Beta coefficients for sentence features in the LRMs, trained on either the crowdsourced votes or the expert votes. The p-values reported are for the crowdsourced model.

4. RESULTS

22 subjects participated in the crowdsourcing study (18 female, average age 23). As quality control for their work, we registered the time they spent on the annotation task. The work of two subjects was disregarded because they completed multiple threads per minute, which would be impossible if they would have taken the task seriously. The median number of items selected in a thread was 7.0 (mean 8.7). The two experts both summarized all 50 threads presented to them. The median number of items selected in a thread was 8.0 (mean 8.0).

We found that for the crowdsourced summaries, the agreement for sentence selection is low: $\kappa = 0.081$. This indicates that the agreement between two raters for selecting sentences from a thread is close to random.[3] The expert agreement is considerably higher: $\kappa = 0.267$, which indicates 'fair agreement'.

We investigated the κ scores of the individual rater pairs in the crowdsourced data to see whether some of the pairs reach the expert agreement level. Figure 1 shows the dispersion of the κ scores for the rater pairs with at least 5 common threads (κ scores are averaged over threads). None of the rater pairs in the crowdsourced data reaches the agreement level of the expert raters. We also investigated the time spent by the individual raters. The mean time spent per thread in the crowdsourced data was 112 seconds while the experts spent 145 seconds on average per thread, which is considerably longer. However, there is no significant correlation between the time spent per thread and the average κ score of an individual rater (Kendall's $\tau = 0.091, p = 0.57$). We also investigated whether relatively more experienced raters reach a higher agreement, but there is no significant correlation between the number of threads summarized and the average κ score of an individual rater (Kendall's $\tau = 0.134, p = 0.41$).

4.1 Characteristics of selected sentences

[3]A κ of zero would indicate an agreement level that is not distinguishable from random selection behavior.

Using the linear regression model (LRM, see Section 3.4) we studied the relation between the sentence features from Table 1 (used as independent variables in the LRM) and the number of times a sentence was selected by the raters (used as dependent variable in the LRM). In the model trained on the crowdsourced data, all features are significant predictors ($p < 0.05$) except for the type-token ratio. In the expert data, the position of the sentence, the relative punctuation count and the author prominence are not significant either. This is most likely caused by the smaller dataset: 50 threads instead of 100.

Figure 2 shows a direct comparison of the two LRMs in terms of the trained β coefficients for the variables in the two models, indicating the feature weights. The figure shows that the coefficients are similar between the models trained on the expert data and the crowdsourced data: in both models, word count (the number of words in the sentence) is the most important sentence characteristic: longer sentences are selected more often than shorter sentences. The position of the post and the position of the sentence have a negative coefficient, which means that posts in the beginning of the thread and sentences in the beginning of the post are more often selected than posts and sentences further down the thread and the post.

We performed a first evaluation of both models by calculating the correlation (Kendall's τ) between the actual number of votes for a post and the number of votes predicted by either of the models in a cross-data setting: we evaluated the expert model on the crowdsourced data and the crowdsourced model on the expert data. The expert model performs a bit better on the crowdsourced data than vice versa: $\tau = 0.304$ for the expert model while $\tau = 0.234$ for the crowdsourced model (both with $p < 0.0001$). As a comparison, the correlations for random models (randomly assigned vote predictions to sentences) are almost 0 and not significant: $\tau = 0.00139, p = 0.94$ for the expert data and $\tau = -0.00428, p = 0.80$ for the crowdsourced data.

5. CONCLUSION

We found that for the task of selecting relevant sentences in patient forum threads, the inter-rater agreement in crowdsourced data is low, close to random ($\kappa = 0.081$). Domain experts achieve a considerably higher, fair, agreement ($\kappa = 0.267$). None of the rater pairs in the crowdsourced

data that have at least five threads in common reach the agreement level of the experts (RQ1). The difference in agreement is not related to time spent on the task or the number of threads summarized by an individual rater.

We trained linear regression models using the two types of reference summaries and a set of sentence features. The models appear to be similar: the characteristics of sentences selected by experts and by crowdsourced raters to be part of the summary are the same: longer sentences are selected more often than shorter sentences and sentences at the beginning of a post and thread are selected more often than sentences later in the post and thread (RQ2).

We conclude that it is possible to train an extractive summarization model on crowdsourced data that is similar to an expert model, even if the inter-rater agreement for the crowdsourced data is low. We speculate that the reasons are: (a) even if two individual raters disagree on the selection of sentences, with five raters per thread the majority opinion is still valuable; (b) the sentence features that we used are generic and robust against noisy selection behavior; (c) the crowdsourced data is bigger than the expert data, which makes it more informative when training the regression model.

One limitation of this work is the low-level nature of the sentence features: we did not include the semantics of the sentences in the model; in future work we will explore more elaborated sentence models. In terms of correlation between the model's predictions and the actual number of votes, the expert model performs a bit better on the crowdsourced data ($\tau = 0.304$) than the crowdsourced model on the expert data ($\tau = 0.234$). In future work, we plan to evaluate the summarization models by running them on unseen forum threads and then showing them to human judges in a blind side-by-side comparison.

With the current data set, we can train robust extractive summarization models by combining the expert and crowdsourced data. This way, we have the reliability of the expert data combined with the size of the crowdsourced data.

Acknowledgments

This work was carried out in the context of the project Discussion Thread Summarization for Mobile Devices (DISCO-SUMO), which is financed by the Netherlands Organisation for Scientific Research (NWO), and the project Patient Forum Miner, which was financed by the SIDN Fund. We thank the volunteers of GIST support international for creating the expert reference summaries.

6. REFERENCES

[1] S. Bhatia, P. Biyani, and P. Mitra. Summarizing online forum discussions–can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131. Association for Computational Linguistics, 2014.

[2] P. Clough, M. Sanderson, J. Tang, T. Gollins, and A. Warner. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17(4):32–38, 2013.

[3] G. Giannakopoulos, J. Kubina, F. Meade, J. M. Conroy, M. Bowie, J. Steinberger, B. Favre, M. Kabadjov, U. Kruschwitz, and M. Poesio. Multiling 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 270, 2015.

[4] P. S. Gill and B. Whisnant. A qualitative assessment of an online support community for ovarian cancer patients. *Patient related outcome measures*, 3:51, 2012.

[5] U. Hahn and I. Mani. The challenges of automatic summarization. *Computer*, 33(11):29–36, 2000.

[6] P.-Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

[7] F. Liu and Y. Liu. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 201–204. Association for Computational Linguistics, 2008.

[8] C. Llewellyn, C. Grover, and J. Oberlander. Summarizing newspaper comments. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pages 599–602, 2014.

[9] M. Marge, S. Banerjee, and A. I. Rudnicky. Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 99–107. Association for Computational Linguistics, 2010.

[10] M. Mitray, A. Singhalz, and C. Buckleyyy. Automatic text summarization by paragraph extraction. *Compare*, 22215(22215):26, 1997.

[11] J. L. Neto, A. A. Freitas, and C. A. Kaestner. Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*, pages 205–215. Springer, 2002.

[12] G. Penn and X. Zhu. A critical reassessment of evaluation baselines for speech summarization. In *ACL*, pages 470–478, 2008.

[13] D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Celebi, D. Liu, and E. Drabek. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 375–382. Association for Computational Linguistics, 2003.

[14] A. S. Tigelaar, R. op den Akker, and D. Hiemstra. Automatic summarisation of discussion fora. *Natural Language Engineering*, 16(02):161–192, 2010.