Lang Resources & Evaluation (2018) 52:461–483
https://doi.org/10.1007/s10579-017-9389-4

ORIGINAL PAPER

Creating a reference data set for the summarization of discussion forum threads

Suzan Verberne[1] · **Emiel Krahmer**[2] · **Iris Hendrickx**[1] · **Sander Wubben**[2] · **Antal van den Bosch**[1]

Published online: 21 April 2017

Abstract In this paper we address extractive summarization of long threads in online discussion fora. We present an elaborate user evaluation study to determine human preferences in forum summarization and to create a reference data set. We showed long threads to ten different raters and asked them to create a summary by selecting the posts that they considered to be the most important for the thread. We study the agreement between human raters on the summarization task, and we show how multiple reference summaries can be combined to develop a successful model for automatic summarization. We found that although the inter-rater agreement for the summarization task was slight to fair, the automatic summarizer obtained reasonable results in terms of precision, recall, and ROUGE. Moreover, when human raters were asked to choose between the summary created by another human and the summary created by our model in a blind side-by-side comparison, they judged the model's summary equal to or better than the human summary in over half of the cases. This shows that even for a summarization task with low inter-rater agreement,

✉ Suzan Verberne
s.verberne@let.ru.nl

Emiel Krahmer
e.j.krahmer@uvt.nl

Iris Hendrickx
i.hendrickx@let.ru.nl

Sander Wubben
s.wubben@uvt.nl

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

[1] Centre for Language Studies, Radboud University, PO Box 9103, 6500 HD Nijmegen, The Netherlands

[2] Department of Communication and Information Sciences, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

a model can be trained that generates sensible summaries. In addition, we investigated the potential for personalized summarization. However, the results for the three raters involved in this experiment were inconclusive. We release the reference summaries as a publicly available dataset.

Keywords Summarization · Discussion forums · Data collection · User study · Inter-rater agreement · Evaluation

1 Introduction

Discussion forums on the web come in many flavors, each covering its own topic and having its own community. The user-generated content on web forums is a valuable source for information. In the case of question answering forums such as StackOverflow and Quora, the opening post is a question and the responses are answers to that question. In these forums, the best answer may be selected by the forum community through voting. On the other hand, in discussion forums where opinions and experiences are shared, there is generally no such thing as 'the best answer'. Moreover, discussion threads on a single topic can easily comprise dozens or hundreds of individual posts, which makes it difficult to find the relevant information in the thread, especially when the forum is accessed on a mobile device.

We address the problem of finding information in long forum threads by automatic summarization. The approach we take in this paper is *extractive* summarization (Hahn and Mani 2000): extracting salient units of text from a source document and then concatenating them to form a shorter version of the document. In most summarization tasks, sentences are used as summarization units. For the summarization of discussion threads it is expected that *posts* are more suitable summarization units than sentences (Bhatia et al. 2014). Therefore, the task that we address in this paper is *post selection*: we aim to identify the most important posts in a discussion. We focus on user evaluation and the creation of a reference data set. We report the results of a user study that we set up to investigate what humans consider to be the most important information in a discussion forum thread, and the results of experiments with an automatic summarizer trained on the collected reference data.

In the literature on thread summarization, a number of features for describing the importance of posts have been identified, such as the position of the post in the thread, the representativeness of the post for the thread, the prominence of the author, the readability of the post, and the popularity of the post (Tigelaar et al. 2010; Bhatia et al. 2014). We hypothesize that the relevance of posts also depends on an *external* variable: the reader of the summary. This hypothesis is motivated by the subjectivity of extractive summarization tasks: It has been shown that if two persons are asked to summarize transcripts of conversations by selecting a subset of utterances from the conversation, their inter-rater agreement is fair at best (Marge et al. 2010; Liu and Liu 2008; Penn and Zhu 2008).

The low inter-rater agreement on the task of extractive summarization has two implications for the development of models for automatic extractive summarization:

First, when using human-labelled data as training data, these data will be inconsistent: information units that are labelled as relevant by one rater are labelled as non-relevant by another rater. Second, the evaluation of extractive summarization systems depends on the individual rater. In this paper we investigate the effect of inter-rater (dis)agreement on the development of an extractive summarizer for long forum threads and we show how multiple reference summaries can be combined to develop a successful model for automatic summarization. We address the following research questions:

RQ1. How useful do human readers consider thread summarization through post selection?
RQ2. What is the desired length of a thread summary?
RQ3. What are the characteristics of the posts that are selected by humans to be included in the summary?
RQ4. How large is the agreement among human raters in selecting posts for the summary?
RQ5. What is the quality of an automatic thread summarizer that is trained on the reference summaries by multiple human raters?

We address these questions with a user study in which reference summaries are created through targeted crowdsourcing among the target group of a large Dutch web forum.[1] We showed long threads to 10 different raters and asked them (a) how useful it would be to have the possibility to see only the most important posts of the thread, and (b) to select the posts that they consider to be the most important for the thread. We analyze their replies to the usefulness question to answer RQ1; we analyze the number of selected posts in order to answer RQ2; we perform a linear regression analysis to answer RQ3; we measure the agreement among the raters in order to answer RQ4. Finally, we show the results of automatic extractive summarization using language-independent features, based on supervised learning and evaluation on human labelled data (RQ5).

Our contributions are the following: (1) we conducted a user study to investigate what are the characteristics of the posts that should be included in the summary, (2) we show that there is only slight to fair agreement between human judges on the task of extractive summarization for forum threads, (3) we present the promising results of automatic extractive summarization using reference summaries by multiple human raters, and (4) we release a dataset of reference summaries for long threads from an open-domain Dutch-language forum.[2]

The remainder of the paper is organized as follows: in Sect. 2, we discuss related work on creating reference summaries for automatic summarization, evaluation metrics for automatic summarization, and methods for discussion thread summarization. In Sect. 3 we describe our methods for data collection, feature extraction and automatic extractive summarization. In Sect. 4 we present our analysis of the

[1] Targeted crowdsourcing is a form of crowdsourcing in which workers are selected who are likely to have the skills needed for the target task, instead of open recruitment on a crowdsourcing platform (Chowdhury et al. 2014, 2015).

[2] The dataset is available from http://discosumo.ruhosting.nl/, under the description "Viva threads with human-assigned votes for post relevance".

human-created summaries and the results for automatic summarization. In the discussion section, Sect. 5, we first answer our research questions, then we investigate the potential for personalized summarization, and we discuss the limitations of the study. Our conclusions are in Sect. 6.

2 Related work

2.1 Creating reference summaries for extractive summarization

For the evaluation of summarization systems, reference summaries created by humans are commonly used. The idea is that by using a reference summary the quality of summarization systems can be compared straightforwardly (Neto et al. 2002). Benchmark reference summaries have been created in the context of the TIPSTER Text Summarization Evaluation Conference (SUMMAC) (Mani et al. 1999, 2002), the NIST Document Understanding Conference (DUC) (Dang 2005) and the NIST Text Analytics Conference (TAC) (Owczarzak and Dang 2011). The explicit focus of DUC 2005 was on the development of evaluation methods that take into account variation in human-created reference summaries. Therefore at least four (and up to nine) different summaries per topic were created, for 50 topics. In the NIST TAC Guided Summarization Task each topic was given to four different NIST assessors.[3]

For creating reference summaries in the case of *abstractive* summarization, raters are typically asked to write a summary of a pre-specified length for a given document or document set. In the context of abstractive discussion thread summarization, a corpus of reference summaries was created by Barker et al. (2016): the SENSEI Annotated Corpus, consisting of reference summaries of user comment threads in on-line news. First, the annotators provided brief (abstract) summaries of each comment in the thread ('labels'). Then these labels were manually grouped under group labels by the annotators, describing the common theme of the group in terms of topic, viewpoints and other aspects. Based on these group labels, the annotators produced a summary for the thread, followed by a final stage in which fragments from the summary were back-linked to messages in the original thread. In the case of *extractive* summarization, which we consider here, a reference summary has the form of a subset of text units selected from the original document (Murray et al. 2005). For most text types, the summarization units are sentences (Gupta and Lehal 2010). In the case of conversation summarization the units are utterances (Marge et al. 2010; Liu and Liu 2008; Murray et al. 2005; Penn and Zhu 2008), and for discussion thread summarization the units typically are posts (Bhatia et al. 2014).

Summarization is an inherently subjective task: not only the length of the created summary differs between human summarizers[4] (Jing et al. 1998), but the content of the summary also differs: raters tend to disagree on the information that should be

[3] http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html.

[4] We will use the word *raters* in the remainder of this paper.

included in the summary. The agreement between two human raters on the content of an extractive summary can be measured using the proportions of selected and non-selected units by the raters, and the percentage of common decisions (selected/non-selected). Agreement is then calculated in terms of Cohen's κ (Radev et al. 2003) for two raters or Fleiss' κ for multiple raters (Landis and Koch 1977). Both have the same general formula, with a different calculation for Pr(e):

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (1)$$

where $Pr(a)$ is the measured agreement (the percentage of common decisions) and $Pr(e)$ is the chance (expected) agreement, based on the proportion of selected and non-selected units by the raters. According to Landis and Koch (1977), the interpretation of κ is as follows: A negative κ indicates structural *dis*agreement. If $\kappa = 0$, there is no agreement between the raters (measured agreement is not higher than chance agreement). A κ between 0.01 and 0.20 indicates slight agreement, between 0.21 and 0.40 indicates fair agreement, between 0.41 and 0.60 indicates moderate agreement, between 0.61 and 0.80 indicates substantial agreement, and between 0.81 and 1.00 indicates (almost) perfect agreement.

For summaries of newswire texts κ scores between 0.20 and 0.50 have been reported (Mitray et al. 1997), with multi-document summaries having a lower inter-rater agreement than single-document summaries (Lin and Hovy 2002). For the summarization of conversation transcripts, the reported κ scores are even lower: between 0.10 and 0.35 (Liu and Liu 2008). One way to address the subjectivity issue is to combine the summaries by multiple raters into one reference model (Jing et al. 1998), for example by using voting over text units: units that are selected by many raters are considered to be more relevant than units selected by few raters (Parthasarathy and Hasan 2015).

To our knowledge, no reference data for extractive summarization of discussion forum threads has been published before. For the Online Forum Summarization (OnForumS) task at MultiLing, no reference summaries were created; the automatically generated summaries were evaluated through crowdsourcing instead (Giannakopoulos et al. 2015; Kabadjov et al. 2015). Thus, compared to previous datasets for extractive summarization, the unique features of our data collection are: (1) the language, genre and domain of the data: Dutch-language discussion forum threads, and (2) the size of the dataset: the larger number of raters per topic (10), combined with a larger number of topics (100) compared to previous data, which allows analysis of the agreement and disagreement between individual raters.

2.2 Evaluation of automatic summarization

The quality of a summary is commonly evaluated using the ROUGE-N evaluation metric, which computes the overlap between the automatic summary and the reference summary in terms of overlapping n-grams (Lin 2004). ROUGE-N is recall-oriented: the number of overlapping n-grams is divided by the number of n-grams in the reference summary. An alternative to ROUGE-N is ROUGE-L, which

is computed as the size of the union of the longest common subsequences ($LCS_{\cup}$) between a sentence in the reference summary and each sentence in the automatic summary, added over all sentences in the reference summary. ROUGE-L has a precision component and a recall component. The recall component divides the sum of $LCS_{\cup}$ by the number of words in the *reference* summary, while the precision component divides the sum of $LCS_{\cup}$ by the number of words in the *automatic* summary. These precision and recall ROUGE-L scores are then combined in the weighted F-measure, where β defines the weight of both components (Lin 2004).

In the case of extractive summarization, the quality of an automatically generated summary can be evaluated using ROUGE (Murray et al. 2005; Tsai et al. 2016), but also using precision, recall, and F1 measures (Neto et al. 2002), considering the selected text units as a whole instead of counting the overlapping content. In this case, precision in is the proportion of text units (sentences, posts) selected by the automatic summarizer that were also included in the reference summary; recall is the proportion of text units included in the reference summary that were also selected by the automatic summarizer; F_1 is the harmonic mean of precision and recall.

Note that use of precision and recall for the evaluation of extractive summarization is a rather strict evaluation method: if the model selects a sentence that was not included in the reference summary, then this sentence is considered a false positive, even if the content of other sentences in the reference summary is largely overlapping with this model-selected sentence. This effect is more severe in the case of texts with much redundancy. ROUGE adopts a more flexible approach to measuring the overlap between the automatic summary and the human reference summary by counting the textual overlap (on the n-gram or word level) between both summaries, instead of making a true/false judgment on the selection of sentences.

2.3 Methods for discussion thread summarization

Most methods for automatic summarization have been developed for domains in which the most important information tends to be located in predictable places, such as scientific articles and news articles (McKeown et al. 2005; Zhou and Hovy 2006). These methods do not work well on texts in which the information is unpredictably spread throughout the text, as we find in internet forums (Wanas et al. 2008).

Over the last decade, some research has been directed at the summarization of forum threads. The oldest work (Zhou and Hovy 2005) focuses on the summarization of technical internet relay chats. In a follow-up paper, the authors argue that forum threads are a form of correspondence, which requires dialogue and conversation analysis (Zhou and Hovy 2006). Tigelaar et al. (2010) take a multi-step approach to thread summarization, involving extensive NLP analysis for feature engineering. They focus on two types of threads: problem solving and discussion, both in nested threads. Central aspects of their method are the detection of the thread structure (responses, quotes and mentions), the prominence of messages, and the prominence of authors in the thread.

An alternative approach to thread summarization is topic modeling (Ren et al. 2011; Llewellyn et al. 2014). In the context of the Online Forum Summarization (OnForumS) task at MultiLing (Giannakopoulos et al. 2015; Kabadjov et al. 2015), Llewellyn et al. (2014) evaluate clustering techniques for summarizing the conversations that occur in the comments section of the UK newspaper the Guardian. They cluster the comments and rank them by their estimated relevance within their cluster. The top comments from each cluster are used to give an overview of that cluster. The authors find that for the task of summarizing newspaper comments topic model clustering gave the best results when compared to a human reference summary. A similar approach is taken by Aker et al. (2016), who address the problem of generating labels for clusters of user comments to online news, as part of the above-mentioned SENSEI project (Barker et al. 2016). They implement a feature-based method using linear regression for optimally combining the features. Their results demonstrate how automatically labeled comment clusters can be used to generate an abstractive summary of a discussion thread.

Bhatia et al. (2014) take a feature-based approach in selecting the most relevant posts from a thread, thereby particularly investigating the use of dialog acts in thread summarization. They evaluate their method on two forums: ubuntuforums.org (problem solving) and tripadvisor.com (experience sharing). Following Bhatia et al. (2014), we approach thread summarization as a post selection problem. Our experimental contribution is that we first conduct an extensive user study to create reference summaries, analyze the collected data, and then build a summarization model based on the combined reference summaries by multiple human raters.

3 Methods

We collected reference summaries through an online user study with target group members of a large open-domain web forum. A reference summary in extractive summarization is defined by the concatenation of the most important bits of information in the text, and hiding or removing the less important fragments in between Hahn and Mani (2000). In order to create these summaries, we presented human raters with a discussion thread and asked them to select the most important posts. In contrast with the other reference data sets, we deliberately did not specify the length of the desired summary and left it to the raters to decide. Each thread was shown to 10 different raters. We analyzed their responses to address our research questions. We trained an automatic extractive summarizer and compared its performance to the reference summaries. In the next sections we discuss each step in detail.

3.1 Data

The Viva Forum[5] is a Dutch web forum with a predominantly female user community. The discussions on the forum are mostly directed at experience and

[5] http://forum.viva.nl.

opinion sharing. Registered users can start new threads and comment on threads. Threads do not contain an explicit hierarchy, and there is no option to 'like' or 'upvote' a post, but users can use the quote option to directly respond to another user's post. The Viva forum has 19 Million page views per month (1.5 Million unique visitors),[6] which makes it one of the largest Dutch-language web forums. We obtained a sample of 10,000 forum threads from the forum owner, Sanoma Media.

The average number of posts in a thread is 33.8 and the median is 7. Around one third (34%) of the threads have more than 10 posts and 21% have at least 20 posts. For our experiment we created a sample of 100 randomly selected threads that have at least 20 posts. Examples of thread titles in our sample are: "Why working out a lot is NOT fun", "Afraid of feelings", "Due date in March 2016" and "Getting married: how to keep the price down".[7] For the purpose of comparison to the literature on thread summarization for problem-solving threads, we added 8 threads from category 'Digi' (comprising technical questions) that have at least 20 posts. Example titles are "help!! water over new macbook air" and "blocked contacts on facebook".

Of threads with more than 50 responses, we only used the first 50 for manual labelling. The median number of posts shown to a rater per thread is 34.

3.2 Manual labelling of sample threads

Through social media and the Radboud University research participation system, we recruited members of the Viva forum target group (Dutch-language, female, aged 18–45) as raters for our study. The users provided some basic information in the login screen, such as how often they have visited the Viva forum in the past month. 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 raters decided themselves how many threads they wanted to summarize. They were paid a gift certificate.

Figure 1 shows a screenshot of the post selection interface. The left column of the screen shows the complete thread; the right column shows an empty table. By clicking on a post in the thread on the left it is added to the column on the right (in the same position); by clicking it in the right column it disappears again. The opening post of the thread was always selected. The raters were given the following instructions in the left column: "Please select the pieces of text (by clicking them one by one) that you think are the most important for the thread. You can determine the number of selected posts yourself." The raters also had the possibility to remove sentences or posts from the selection by clicking the selected items. The instruction text in the right column reads: "By reading your selection of posts you can check whether you created a good summary of the topic. You can remove posts from your selection by clicking on them. Click on the 'Submit selection' button if your selection is final. If you did not select any posts, please explain in the comments field why." We intentionally did not pre-require a specific number of posts to be

[6] http://www.sanoma.nl/merken/bereik/viva/.

[7] Translated to English for the reader's convenience.

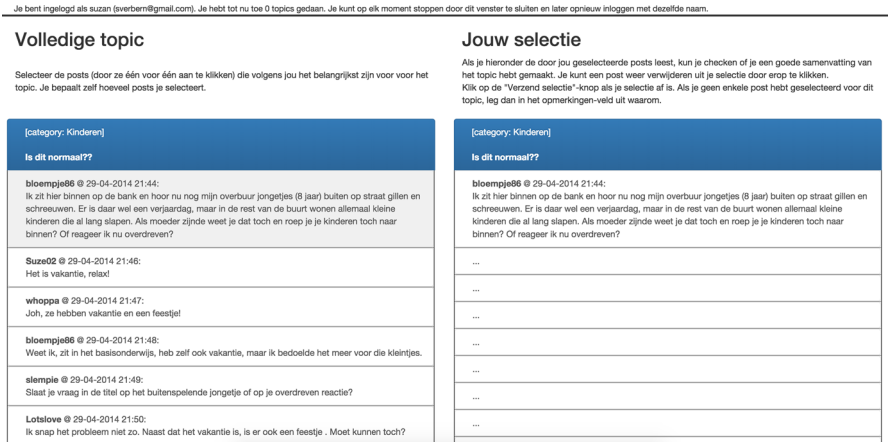

Fig. 1 A screenshot of the post selection interface. The *top-most line* provides information on the number of threads that the rater has summarized. The *left column* ('Volledige topic') shows the full thread while the *right column* ('Jouw selectie') shows the rater's selected threads (with the first post always selected). In the *blue header*, the category and title of the thread are given. Each cell is one post, starting with the author name and the timestamp

selected for the summary because we wanted to investigate what the desired summary size was for the raters.

We also asked the raters to indicate their *familiarity* with the topic of the thread (scale 1–5, where 1 means 'not familiar at all' and 5 means 'highly familiar') and how *useful* it would be for this thread to have the possibility to see only the most important posts (scale 1–5). In case they chose a usefulness score of 1, they were asked to choose between either of the options 'none of the posts are relevant', 'all posts are equally relevant' or 'other reason'. We gave room for additional comments.

3.3 Feature extraction

In order to answer the question "What are the characteristics of the posts that are selected by humans to be included in the summary?" (RQ3), we investigated the relationship between post features and the selection of posts using a linear regression analysis. With 10 raters per thread, each post receives between 0 and 10 votes. We used the number of votes for a post as the dependent variable in the regression analysis. We argue that the number of votes for a post is an indicator of its relevance: a post that is selected by all 10 raters can be expected to be more relevant than a post that is selected by only one or two raters. The post features that we used as independent variables are taken from the literature on extractive summarization (Weimer et al. 2007; Tigelaar et al. 2010; Bhatia et al. 2014). The features are listed in Table 1. All features are language-independent.

Table 1 Post features used as independent variables in the regression analysis for answering the question "What are the characteristics of the posts that are selected by humans to be included in the summary?"

Category	Description
Position	Absolute position in the thread
Position	Relative position in the thread
Popularity	# of responses (quotes) to the post
Representativeness	Cosine sim between post and thread (tf-idf weighted term vectors)
Representativeness	Cosine sim between post and title (tf-idf weighted term vectors)
Readability	Word count
Readability	Unique word count
Readability	Type-token ratio
Readability	Relative punctuation count
Readability	Average word length (# of characters)
Readability	Average sentence length (# of words)
Author prominence	Proportion of posts in thread by author of current post

3.4 Automatic extractive summarization

Creating an extractive summary using human-created example summaries, is inherently a binary classification problem, where each post is classified into one of two classes: selected or not selected (correct or incorrect; relevant or irrelevant) (Neto et al. 2002). Having reference summaries created by 10 different raters allowed us to consider the selection of posts as a graded relevance problem: the more raters selected a post, the more relevant it is. Therefore, we approached the extractive summarization task as a regression task, where the relevance of the post is represented by the number of votes it received.

For evaluation purposes, we randomly divided the data in 5 partitions while keeping all posts belonging to the same thread together in the same partition. In a fivefold cross validation setup, we use 3 partitions for training, 1 for tuning and 1 for testing. We perform a linear regression analysis on the training data, using the number of votes as dependent variable and the features listed in Table 1 as independent variables. After tuning the threshold parameter (see Sect. 4.5) on the tuning set, we use the model to predict the number of votes for each of the posts in the test set. Applying the tuned threshold on this predicted value results in the selection of posts.

4 Results

57 raters participated in the study: all female, average age 27 (median 25, SD 7.7, min 18, max 44). We disregarded two threads because long URLs in some of the posts caused the display in our annotation interface to be disturbed, so 106 threads remain. All were summarized by 10 raters. Table 2 illustrates what the raw collected data looks like.

Table 2 Example of the collected data: the set of selected posts (represented by their post id, '1' being the first comment after the opening post) by the 10 raters for one example thread. Rater 6 deselected post no 14 again, after first selecting it

Rater #	Selected posts
1	1 4 5 11
2	3 4 6 7 14 22 29
3	4 5 11 13 19 20 21 27 28
4	4 14 28
5	1 4 6 13 15 18 19 21 27 28
6	1 4 9 11 13 14 15 −14 19 20 27 28
7	1 3 4 8 9 11 12 13 18 19 21 22
8	1 8 13 19 27
9	1 4 8 9 13 18 27 28
10	1 4 5 9 11 13 14

The majority of threads (58%) was summarized by raters who indicated that they visited the Viva forum once or twice in the past month. The correlation between forum visit frequency and familiarity with the topic is very weak (Kendall's $\tau = 0.08$, $p = .006$), indicating that the raters who did not visit the forum were equally familiar with the topics as the raters who did visit the forum.

4.1 Usefulness of thread summarization (RQ1)

The median usefulness score over all threads is 3 (on a 5-point scale) with a standard deviation of 1.14 (averaged over threads). For 92% of the threads, at least one rater gave a usefulness score of 3 or higher and for 62% of the threads, at least half of the raters gave a usefulness score of 3 or higher. We manually analyzed the comments that were posted by the raters in the optional comments field. All raters together posted 155 comments (15% of summarized threads), mostly when they assigned a usefulness score of 1. For 29% of the threads, at least half of the raters gave a usefulness score of 1. For these cases with very low usefulness, the raters indicated that either all posts are equally important (typically threads in which opinions are shared) or none of the posts are important ('chatter threads').

We investigated whether the usefulness of thread summarization can be predicted from characteristics of the thread. We performed a linear regression analysis with the median usefulness score as dependent variable and the following *thread features* as independent variables:[8]

- number of posts in the thread
- length of the title
- length of the opening post
- average post length

[8] Note that these variables are defined on the level of the full thread, whereas the variables in Table 1 were defined on the level of a single post.

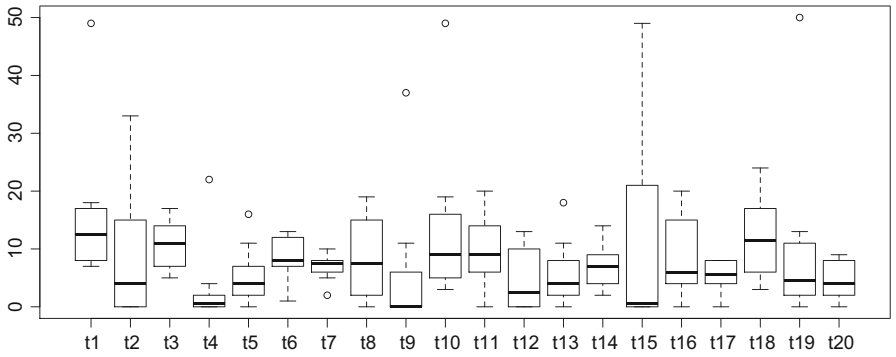

Fig. 2 Box plot showing the dispersion of the number of selected posts by the 10 raters for the first 20 threads (t1–t20)

- number of question marks in the opening post
- the average cosine similarity between each of the posts and the complete thread
- the average cosine similarity between each of the posts and its previous post
- the forum category.

We found that none of the numeric variables is a significant predictor of the usefulness of summarization of a thread (p values are all above 0.1). The only significant predictor for the median usefulness score is the category value 'Digi' ($p = .0296$)—the technical question category. The mean usefulness score for the 'Digi' threads was 3.3 (n = 13) while the mean usefulness score for all other threads was 2.4 (n = 93). A Welch t test for independent samples shows that this difference is significant ($p = .0029$). Thus, we observe evidence that summarization of technical issues is considered to be more useful than summarization of discussions directed at opinion and experience sharing, but the usefulness of thread summarization cannot be predicted from numeric characteristics of the thread.

4.2 Desired length of a thread summary (RQ2)

Only a small proportion of posts were selected by all raters: 0.5%. Almost one in five posts (19.1%) was selected by at least five raters. A quarter of the posts (24.5%) were selected by none of the raters. This indicates that the raters are more often unanimous about a post being *irrelevant than about a post being* relevant.

The median number of posts selected in a thread by the human raters was 7 (mean 8.9). The standard deviation over raters was high: 6.4 (averaged over threads), the minimum was 0 and the maximum was 50. Figure 2 shows the dispersion of the number of selected posts by the 10 raters for the first 20 threads in our dataset, as an illustration of the large variance in the number of selected posts.

We investigated the effect of two variables on the number of selected posts: the thread and the rater. An ANOVA shows that both have a statistically significant effect, but the effect of the rater is larger: for the rater, $F(56, 1009) = 15.7, p < .001, \eta^2 = 0.47$ while

Table 3 Post features that are significant predictors for the number of selected votes, sorted by the absolute value of the regression coefficient β; the independent variable with the largest effect (either positive or negative) is on top of the list

Category	Feature	β coef
Position	Absolute position in the thread	−0.78***
Representativeness	Cosine sim between post and thread	0.52***
Readability	Unique word count	0.37***
Readability	Type-token ratio	−0.22***
Readability	Average word length (# of chars)	0.22***
Author prominence	Proportion of posts in thread by author of post	−0.15***
Readability	Relative punctuation count	−0.15***
Representativeness	Cosine sim between post and title	0.11***
Popularity	# of responses (quotes) to the post	−0.08**

** $p < .01$; *** $p < .001$

for the thread, $F(105, 954) = 2.51, p < .001, \eta^2 = 0.22$. In addition, we found that the correlation between the number of selected posts and the total number of posts in the thread is very weak (Spearman's $\rho = 0.101$), indicating that the desired length of the summary does not depend on the length of the thread. This indicates that the desired length of a thread summary is personal, depending on the reader of the thread.

4.3 Characteristics of the selected posts (RQ3)

Table 3 shows a ranking of the post features that significantly predict the number of votes for a post, according to the linear regression analysis. The table shows that almost all features have a significant effect, although some effects are small. The top-ranked features are the Position, Representativeness and Readability features. Representativeness is known to be an important feature for extractive summarization (sometimes referred to as *centrality)* (Erkan and Radev 2004). The negative coefficient for the absolute position feature indicates that posts in the beginning of the thread tend to get more votes than posts further down the thread. This is in line with the work by Bhatia et al. (2012), in which it was found that the position of the post in the thread is one of the most important features for determining the purpose of the post.

The negative coefficient of type-token ratio—which is a measure of lexical diversity—seems surprising: we had expected that a higher type-token ratio would lead to more votes. The effect can be explained by the negative correlation between type-token ratio and post length, which has been reported in the literature before (Richards 1987).

4.4 Agreement between human raters (RQ4)

We investigated the agreement between human raters on which posts should be included in the summary. In calculating the agreement, we could either use Fleiss' κ

for multi-rater data, or Cohen's κ for separate pairs of raters. The former seems more appropriate as we have 10 raters per thread, but the latter would be conceptually correct because we do not have *the same 10 raters* for each thread. We therefore report both Fleiss' κ and Cohen's κ.

In order to calculate Cohen's κ in a pairwise fashion, we computed the agreement between each pair of raters for each thread (all possible unique pairs of 10 raters per thread: 45 pairs). If both raters selected 0 posts, we set $\kappa = 1$. We measured the agreement for each pair and then computed the mean over all pairs and over all threads. In addition, we report the Jaccard similarity coefficient, which is calculated as the size of the intersection divided by the size of the union of the two sets of selected posts, If both raters selected 0 posts, we set Jaccard = 1.

We found that Fleiss' $\kappa = 0.219$ for our data, which indicates fair agreement. We found a mean Cohen's κ of 0.117, which indicates a slight agreement. As a comparison, Liu and Liu (2008) reported κ scores between 0.11 and 0.35 for utterance selection in summarizing conversations. We found that the mean Jaccard coefficient over human–human pairs for all threads was 0.259.

4.5 Results of the automatic summarization (RQ5)

In this section, we evaluate our automatic summarizer in two ways: First, we measured the overlap between the posts selected by the model and the posts selected by each of the human summarizers, in terms of Jaccard coefficient, Cohen's κ, precision, recall, F1, and ROUGE (Sect. 4.5.1). Second, we had raters evaluate the summaries in a blind pairwise comparison between their own summary, the summary of another human summarizer and the model's summary (Sect. 4.5.2).

4.5.1 Evaluation of the summarization model against human reference data

We trained a linear regression model for extractive summarization on the basis of the post features in Table 1. Again, the dependent variable is the number of votes for a post by the human raters. Although the agreement between the raters on the selection of posts was slight, our assumption is that the number of votes for a post is a measure for its relevance.

In order to perform extractive summarization for unseen threads, we have to decide on the number of posts that are included in the summary. The median number of selected posts over all threads and all raters was 7, but the divergence between threads and raters was large. Thus, fixing the number of selected posts over all threads would not lead to good summaries for all readers. Instead, we set a threshold on the outcome of the regression model (predicted number of votes for each post) that, when applied to all posts, leads to a median of 7 posts selected per thread. In the fivefold cross validation setup, we tuned this threshold on the held-out tuning set for each training set.

For the five tuning sets, we found thresholds between 3.35 and 3.94, each leading to a median of 7 selected posts for the threads in the tune set. Over all threads in all test partitions, the median number of selected posts was also 7, but the deviation is

smaller than in the human-summarized data (mean 7.2, SD 4.9, max 21, min 0—in the human-summarized data, SD was 6.4, max 50 and min 0.). For 95% of the threads, the model selects at least one post.

We implemented three baselines to compare our model with:

- Random baseline: selects 7 posts randomly for each thread
- Position baseline: selects the first 7 posts of each thread
- Length baseline: selects the 7 longest posts of each thread.

The latter two baselines are informed baselines, given the finding that the position and the length of a post are important indicators for the post's relevance.

We calculated the Jaccard index and Cohen's κ between the model and each human rater, and compared these to the mean Jaccard and κ scores for agreement between two human raters. In addition we calculated precision, recall, F-score and ROUGE per rater for each thread,[9] and report mean results over all threads and all raters.

As introduced in Sect. 2.2, ROUGE traditionally is a recall-oriented measure (Lin 2004). One implication is that ROUGE-N is higher for longer summaries (because the longer the automatic summary, the larger the overlapping set), which is the reason that in most evaluation tasks, the length of the summary is pre-defined. In our data however, the length of the summary was not pre-defined and although all baselines generate the same number of posts, the concatenated length of the selected posts is structurally longer for the length baseline than for the other settings. We therefore also report ROUGE-L F_1, the harmonic mean of ROUGE-L precision and ROUGE-L recall. We calculated these as follows (see also Sect. 2.2; Lin (2004)):

$$Recall_{lcs} = \frac{\sum_{i=1}^{u} |LCS_{\cup}(r_i, A)|}{m} \tag{2}$$

$$Precision_{lcs} = \frac{\sum_{i=1}^{u} |LCS_{\cup}(r_i, A|)}{n} \tag{3}$$

Here r_i is a post from the reference summary, u is the total number of posts in the reference summary, A is the set of posts in the automatic summary, $LCS_{\cup}(r_i, A)$ is the union of the longest common subsequences between ri and each post in A, and $|LCS_{\cup}(r_i, A)|$ is the size of this set. m is the total number of words in the reference summary (summed over all r_i) and n is the total number of words in the automatic summary (summed over all $a_i \in A$).

The results are in Table 4. The results show that we obtained almost equal agreement scores for the human–model pairs as for the human–human pairs: mean Jaccard was 0.271 and mean κ was 0.138, indicating that our model on average has a slightly higher agreement with human raters than human raters themselves. We found a mean precision over all raters and threads of 44.6%, a mean recall of 45.9% and a mean F_1 of 45.2%, thereby outperforming all baselines by a large margin.

[9] If both the model and the human rater selected 0 posts, we set Precision = Recall = ROUGE = 1.

Table 4 Results of our summarization model, compared to 3 baselines and human–human agreement

Human–human comparison				
Mean Jaccard				0.259
Mean Kappa				0.117
	Rand baseline	Position baseline	Length baseline	Our model
Human–model comparison				
Mean Jaccard	0.121	0.204	0.224	**0.271**
Mean Kappa	−0.085	0.06	0.092	**0.138**
Mean Precision	25.9%	37.1%	39.1%	**44.6**%
Mean Recall	20.4%	34.7%	37.3%	**45.9**%
Mean F1	22.8%	35.9%	38.2%	**45.2**%
Mean ROUGE-1	42.5%	43.6%	**72.1**%	68.0%
Mean ROUGE-2	24.2%	31.1%	**60.3**%	60.0%
Mean ROUGE-L F_1	6.5%	8.7%	11.7%	**12.4**%

Precision, recall, F1 and ROUGE are macro averages over the threads. Boldface indicates the best performing model according to each evaluation metric

The results with ROUGE show a different pattern. First, we see high scores for ROUGE-1, comparable to the ROUGE-1 scores found in related work on extractive summarization of speech transcripts (Murray et al. 2005). We also observe that our model is outperformed by the length baseline in terms of ROUGE-1 and ROUGE-2, which is most likely caused by the recall-oriented nature of the metrics. In terms of ROUGE-L, we observe relatively low scores (12.4% for our model) compared to scores reported in the literature: Murray et al. (2005) report ROUGE-L scores between 20 and 30%, and Wong et al. (2008) around 31%. This is probably because of the length of the summarization unit: posts are longer than sentences (59 words on average) and longest common subsequences are relatively short for two non-identical posts. Our model does outperform the length baseline by a small margin. According to a paired t test on the ROUGE-L scores for individual thread–rater pairs ($n = 1060$) the difference is significant on the 0.05-level ($p = 0.046$).

4.5.2 Evaluation using pairwise human judgments

Figure 3 exemplifies two summaries of the same thread, one created by our model and one created by one of the human raters. The example illustrates that thread summarization is subjective: The reader's opinions and beliefs co-define what is important. The example also illustrates that two different summaries can still both be good summaries. This leads us to believe that it is possible that readers are satisfied by a summary, even though the summary is different from the summary that they would have created themselves. We investigated this in a pairwise judgment study, in which human raters were presented with two summaries for the same thread and were asked to judge which of the two is better. In the user interface,

prisje94: My boyfriend's ex keeps stalking him and she has been provoking responses from him over 2 months now. He had removed her on facebook whatsapp in his phone etc of which we thought it would help. [...] But when I logged in on his facebook account today she suddenly got unblocked. Do you think this could be an error caused by facebook or would she have hacked his account? [...]	
Summary created by our model	**Summary created by one of the human raters**
... 3 posts skipped ...	... 2 posts skipped ...
	PlofKipje84: Facebook does not unblock by itself.
	coco95: That is not possible he did it himself and is lying about it. If you unblock someone you can only block them again after 48 hours.
sinnombre: Don't login to his account anymore. Unless he asked you to. Maybe he unblocked her, because he secretly thinks her messages are exciting and fun. [...]	**sinnombre**: Don't login to his account anymore. Unless he asked you to. Maybe he unblocked her, because he secretly thinks her messages are exciting and fun. [...]
... 2 posts skipped ...	... 2 posts skipped ...
10012015anoniem: The unblocking is not an error by facebook but a conscious action by your boyfriend. He is lying to you, it's up to you to have an opinion about it. And up to him to have an opinion about you hacking his facebook. Nice relation it seems, both are not to be trusted! [:facepalm:]	**10012015anoniem**: The unblocking is not an error by facebook but a conscious action by your boyfriend. He is lying to you, it's up to you to have an opinion about it. And up to him to have an opinion about you hacking his facebook. Nice relation it seems, both are not to be trusted! [:facepalm:]
... 4 posts skipped ...	... 5 posts skipped ...
NYC: Couldn't it be an error by facebook? I don't have experienced with blocked contacts but I've had it with other settings. [...]	
... 3 posts skipped ...	... 2 posts skipped ...
	aarinda: I think you should ask this to your boyfriend. Not to us. I haven't blocked people on FB myself but I have experiences with a spontaneous reset (multiple times) of my privacy settings, just like a few people above. But go talk to your boyfriend. [...]
prisje94: When I logged into his account he was sitting next to me and was OK with it. I discovered that she was unblocked because we could see responses of her again that she had posted to his photos before. You shouldn't judge so soon that I don't trust him. [...]	
... 5 posts skipped ...	... 6 posts skipped ...

Fig. 3 Two summaries of the same thread. The *top row* shows the opening post of the thread (which is always included in the summary). The *left column* shows the summary created by our model; the *right column* shows the summary created by one of the human raters. The selected posts are shown, the unselected posts are hidden ('skipped'). The posts have been translated to English for the reader's convenience, and long posts have been cropped to fewer sentences in order to save space. The author names are printed in boldface

the posts that were not included in the summary were hidden behind a button with the text 'expand response'. The judgments were made on a 5-point scale ranging from 'the summary on the left is much better' to: 'the summary on the right is much better', and an additional 'don't know' option with room for comments.

The raters who participated in this evaluation were the four raters who summarized the most threads in the first study. We selected from our data all threads that were summarized by at least two of them, and we kept only the threads for which all raters gave a usefulness score of 3 or higher. This led to a subset of 52 threads, of which the four raters had respectively summarized 41, 39, 31, and 32 threads in the first session. In the evaluation session the raters were presented with three summaries of the threads they had summarized before: their own summary, the summary by one of the other raters, and the summary generated by our model. Two of these summaries were shown side-by-side, resulting in 3 pairs per thread.

Table 5 The results of the blind side-by-side comparisons by human judges

Overall results	% wins
Rater's own summary	38.9
Summary by another human rater	25.9
Summary by our model	19.3
Tie	15.9
Direct comparisons between 'other' and 'model'	**% wins**
Summary by another human rater	48.3
Summary by our model	42.5
Tie	9.2

The pairs of summaries and the threads were presented in random order. In total, the raters judged $3 * (41 + 39 + 31 + 32) = 429$ pairs.

The results are in the upper half of Table 5. The rater's own summary clearly wins the most comparisons, which is interesting because there was more than a month between the first study in which the summaries were created, and the second study in which the summaries were judged. This indicates that the task is indeed subjective: post selection depends on personal preferences.

In addition, we addressed the question: did the raters judge the summaries by another human ('other') better than the automatically generated summaries ('model')? The upper half of Table 5 suggests that the human summaries received more votes than the model-created summaries, but this includes the comparisons to the rater's own summary. We therefore also analyzed the 143 direct comparisons between 'other' and 'model'. In these direct comparisons, the model was judged as equal to or better than the human summary in 51.7% of the cases. If we disregard the ties, the model won 42.5% of the comparison. According to a z-test comparing the distributions of 'model' and 'other' in this sample to a sample of 10,000 random votes for 'model' or 'other', this distribution is significantly different from random ($z = 1.67, p = .047$). Thus, the human-generated summaries did indeed receive more votes than the model-generated summaries, although the difference is small.

5 Discussion

5.1 Answers to the research questions

RQ1. How useful do human readers consider thread summarization through post selection? We found that thread summarization through post selection is considered to be useful for the majority of long threads. We cannot predict on the basis of quantitative thread features whether summarization of a thread will be useful. Therefore, the best strategy for an online summarizer seems to create a summary for all long threads. Furthermore, we found that summarization of threads addressing

technical issues is regarded more useful than summarization of discussions directed at opinion and experience sharing.

RQ2. What is the desired length of a thread summary? We found that the median number of posts in a human-created summary is 7, but that the desired length of a thread summary shows a large deviation between human raters. The length should therefore be tunable by the reader. This can in practice be implemented using a slider in the interface with the text "show more/fewer posts".

RQ3. What are the characteristics of the posts that are selected by humans to be included in the summary? We found that representativeness, readability and position of posts are the most important criteria for extractive thread summarization.

RQ4. How large is the agreement among human raters in selecting posts for the summary? We found that human raters tend to disagree on which posts should be included in the thread summary: mean Cohen's κ between raters is 0.117, which indicates a slight agreement and Fleiss' $\kappa = 0.219$, which indicates fair agreement.

RQ5. What is the quality of an automatic thread summarizer that is trained on the reference summaries by multiple human raters? We found that a model trained on the human-labelled data obtained a reasonable precision and recall (44.6% and 45.9% respectively), outperforming random and informed baselines by a large margin. ROUGE-L scores for the model are less convincing (12.4%) but still significantly better than the best baseline. The agreement between the model and the human raters was comparable to the mean agreement between human raters: $\kappa = 0.138$. In a pairwise judgment experiment, we found that when human raters were asked to choose between the summary created by another human and the summary created by our model in a blind side-by-side comparison, they picked the model's summary in 42.5% of the cases.

5.2 Potential for personalized summarization

Because of the low inter-rater agreement on post selection, we investigated the potential of personalization for the task. For training personalized models we used the three raters (A, B and C) with the largest number of reference summaries: 90, 72 and 70 respectively. On these personal data sets, we trained and evaluated personal post selectors, using 3 partitions from the rater's own data for training, 1 for tuning and 1 for testing. We also evaluated the personal post selectors on the data from the other two raters. In this cross-user setting, we trained the model on four partitions of the training rater, tuned the threshold on the remaining partition of the training rater, and evaluated on the complete data for the test rater. We also evaluated the generic model for each of the three test raters. The results are shown in Table 6. For each test rater, the result for the best performing model is printed in boldface.

The results are inconclusive: for rater C, the personalized model gives markedly better results than the generic model. For rater A, the personalized model seems slightly better than the generic model, but this difference is not significant according to a paired t test on the obtained F-scores per thread ($t = 1.3165$, $p = .191$). The same holds for rater B, where the generic model seems better, but the differences between the F-scores are not significant ($t = -1.772$, $p = .0808$). Thus, we cannot conclude on the basis of these results that a personalized model for extractive

Table 6 F_1 scores obtained for the extractive summarization by three individual raters (A, B and C) using four different models. The scores on the diagonal were obtained through cross validation on the data by the test rater

Trained on →	A (%)	B (%)	C (%)	Generic (%)
F_1 for A	**43.4**	41.8	38.7	41.3
F_1 for B	43.8	39.4	41.9	**44.3**
F_1 for C	49.2	46.8	**59.9**	48.7

summarization is better than a generic model; this might be an interesting direction for future research.

5.3 Limitations of the study

After analyzing the results of our study, we identified three limitations: First, in the instructions for the raters we asked them to select the most important posts without asking them to be *concise*. In case of redundancy between relevant posts, some of the raters tended to select all these relevant posts ("All posts seem equally important" was sometimes mentioned in the comments field). Second, the inter-rater agreement metrics such as κ are limited in the sense that agreement is measured in a discrete way: If the raters select two distinct but highly similar posts, this is counted as a non-match. Third, our extractive summarization method selects posts independently of each other, without considering the relations between posts (except for the number of quotes, used to estimate the popularity feature). For some threads, the model does not select any posts because none of them gets a predicted number of votes above the threshold, while for other threads, the model selects (almost) all posts to be included in the summary, thereby creating redundancy. Both issues should be addressed in future work.

6 Conclusions

We studied the (dis)agreement between human judges on an extractive summarization task: post selection for long discussion forum threads. In a user study we found that for the majority of long threads on an open-domain discussion forum, raters value the idea of thread summarization through post selection. However, when they were asked to perform the extractive summarization task, the inter-rater agreement was slight to fair.

We trained and evaluated a generic model for extractive summarization by combining the reference summaries created by the 10 raters. The model performs similar to a human rater: the agreement between the model and human raters is not lower than the agreement among human raters. Moreover, in a side-by-side comparison between a summary created by our model and a summary created by a human rater, the model-generated summary was judged as equal to or better than the human summary in 51.7% of the cases. The raters had a preference for their own

summary, although they created it a month earlier. This indicates that personal preferences do play a role in extractive summarization, although our experiments with personalized models were inconclusive—perhaps due to data sparseness. Our results indicate that we can generate sensible summaries of long threads on an open-domain discussion forum.

In the near future, we plan to (1) extend our work to other forum types and languages, such as Facebook discussion threads; (2) train and evaluate pairwise preference classifiers for extractive summarization; (3) implement and evaluate sentence-level summarization of long threads; and (4) experiment with query-based summarization using the query logs of the Viva forum.

Our collection of human reference summaries is unique in terms of domain (discussion forum threads) and size (106 threads by 10 raters). We release the reference summaries as a publicly available dataset.

References

Aker, A., Paramita, M., Kurtic, E., Funk, A., Barker, E., Hepple, M., et al. (2016). Automatic label generation for news comment clusters. In *The 9th international natural language generation conference*, p. 61.

Barker, E., Paramita, M., Aker, A., Kurtic, E., Hepple, M., & Gaizauskas, R. (2016). The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pp. 42–52.

Bhatia, S., Biyani, P., & Mitra, P. (2012). Classifying user messages for managing web forum data. In *Fifteenth international workshop on the Web and Databases (WebDB 2012)*, Citeseer.

Bhatia, S., Biyani, P., & Mitra, P. (2014). 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)* (pp. 2127–2131). Association for Computational Linguistics.

Chowdhury, S. A., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., & Klasinas, I. (2014). Cross-language transfer of semantic annotation via targeted crowdsourcing. In *INTERSPEECH*, pp. 2108–2112.

Chowdhury, S. A., Calvo, M., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., et al. (2015). Selection and aggregation techniques for crowdsourced semantic annotation task. In *Sixteenth annual conference of the international speech communication association*.

Dang, H. T. (2005). Overview of DUC 2005. In *Proceedings of the document understanding conference*, pp. 1–12.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Giannakopoulos, G., Kubina, J., Meade, F., Conroy, J. M., Bowie, M., Steinberger, J., et al. (2015). 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*, p. 270.

Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, *2*(3), 258–268.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. *Computer*, *33*(11), 29–36.

Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In *AAAI symposium on intelligent summarization*, pp. 51–59.

Kabadjov, M., Steinberger, J., Barker, E., Kruschwitz, U., & Poesio, M. (2015). Onforums: The shared task on online forum summarisation at multiling'15. In *Proceedings of the 7th forum for information retrieval evaluation* (pp. 21–26). ACM.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *1*, 159–174.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pp. 74–81.

Lin, C. Y., & Hovy, E. (2002). Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 workshop on automatic summarization* (Vol. 4, pp. 45–51). Association for Computational Linguistics.

Liu, F., & Liu, Y. (2008). 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* (pp. 201–204). Association for Computational Linguistics.

Llewellyn, C., Grover, C., & Oberlander, J. (2014). Summarizing newspaper comments. In *Proceedings of the eighth international AAAI conference on weblogs and social media*, pp. 599–602.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., & Sundheim, B. (1999). The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of the ninth conference on European chapter of the association for computational linguistics* (pp. 77–85). Association for Computational Linguistics.

Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., & Sundheim, B. (2002). SUMMAC: A text summarization evaluation. *Natural Language Engineering*, *8*(01), 43–68.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). 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* (pp. 99–107). Association for Computational Linguistics.

McKeown, K., Passonneau, R. J., Elson, D. K., Nenkova, A., & Hirschberg, J. (2005). Do summaries help? In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 210–217). ACM

Mitray, M., Singhalz, A., & Buckleyyy, C. (1997). Automatic text summarization by paragraph extraction. *Compare*, *22215*(22215), 26.

Murray, G., Renals, S., & Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proceedings of INTERSPEECH 2005—EUROSPEECH*.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002). Automatic text summarization using a machine learning approach. In *Advances in artificial intelligence* (pp. 205–215). Berlin: Springer.

Owczarzak, K., & Dang, H. T. (2011). Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *Proceedings of the text analysis conference (TAC 2011)*. Gaithersburg, Maryland, USA.

Parthasarathy, S., & Hasan, T. (2015). Automatic broadcast news summarization via rank classifiers and crowdsourced annotation. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5256–5260). IEEE

Penn, G., & Zhu, X. (2008). A critical reassessment of evaluation baselines for speech summarization. In *ACL*, pp. 470–478.

Radev, D. R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., et al. (2003). Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st annual meeting on association for computational linguistics* (Vol. 1, pp. 375–382). Association for Computational Linguistics.

Ren, Z., Ma, J., Wang, S., & Liu, Y. (2011). Summarizing web forum threads based on a latent topic propagation process. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 879–884). ACM.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language*, *14*(02), 201–209.

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

Tsai, C. I., Hung, H. T., Chen, K. Y., & Chen, B. (2016). Extractive speech summarization leveraging convolutional neural network techniques. In *Proceedings of 2016 IEEE workshop on spoken language technology*.

Wanas, N., El-Saban, M., Ashour, H., & Ammar, W. (2008). Automatic scoring of online discussion posts. In *Proceedings of the 2nd ACM workshop on information credibility on the Web* (pp. 19–26). ACM.

Weimer, M., Gurevych, I., & Mühlhäuser, M. (2007). Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 125–128). Association for Computational Linguistics.

Wong, K. F., Wu, M., & Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics* (Vol. 1, pp. 985–992). Association for Computational Linguistics.

Zhou, L., & Hovy, E. (2005). Digesting virtual geek culture: The summarization of technical internet relay chats. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 298–305.

Zhou, L., & Hovy, E. H. (2006). On the summarization of dynamically introduced information: Online discussions and blogs. In *AAAI Spring symposium: Computational approaches to analyzing weblogs*, pp. 237–246.