Towards Opinion Summarization from Online Forums

Ying Ding
School of Information Systems
Singapore Management University
ying.ding.2011@smu.edu.sg

Jing Jiang
School of Information Systems
Singapore Management University
jingjiang@smu.edu.sg

Abstract

Summarizing opinions expressed in online forums can potentially benefit many people. However, special characteristics of this problem may require changes to standard text summarization techniques. In this work, we present our initial attempt at extractive summarization of opinionated online forum threads. Given the nature of user generated content in online discussion forums, we hypothesize that besides relevance, text quality and subjectivity also play important roles in deciding which sentences are good summary sentences. We therefore construct an annotated corpus to facilitate our study of extractive summarization of online discussion forums. We define a set of features to capture relevance, text quality and subjectivity, and empirically test their usefulness in choosing summary sentences. Using unpaired Student's t-test, we find that sentence length and number of sentiment words have high correlations with good summary sentences. Finally we propose some simple modifications to a standard Integer Linear Programming based summarization framework to incorporate these features.

1 Introduction

With the growing popularity of social media, people often share their experience and opinions openly on the Internet. Especially when a controversial event happens, there are many different opinions expressed in online forum threads, including judgement of the people and organizations involved in the event and suggestions for future changes. Since it is too time consuming to go through all the posts of a thread to understand every individual's opinion, summarizing online discussion forums becomes an important task that may benefit people including government policy makers and social scientists. While text summarization has been extensively studied, summarizing noisy and subjective user-generated content is still an under-explored area. A vast body of work has been done on summarizing online product reviews, but because of the special properties of product reviews, opinion summarization of product reviews tends to focus on product aspect identification and sentiment polarity classification. When it comes to summarizing general online discussions, particularly discussions on controversial topics such as a policy or a social issue, the challenges we face can be very different from summarizing product reviews.

Table 1 shows a set of summary sentences selected by a state-of-the-art summarization method (Gillick and Favre, 2009) from a forum thread on criticizing parliament ministers sleeping in a meeting. We can see that the summary contains low-quality sentences and some sentences do not express opinions. This result shows that traditional text summarization techniques, which only consider text representativeness, may not be able to summarize opinions from online forums very well.

In particular, we hypothesize that two important factors should be considered for summarizing online discussions. First, because forum posts are often noisy, with misspelling, broken sentences and online jargon, text quality should be considered for selecting good candidate summary sentences. Second, because the goal of summarizing discussion forums is mainly to capture online users' opinions, there should be a preference to choose subjective sentences for summaries.

To test the hypotheses above, we need ground truth summaries. Unfortunately, to the best of our knowledge we are not aware of existing bench-

Proceedings of Recent Advances in Natural Language Processing, pages 138–146,
Hissar, Bulgaria, Sep 7–9 2015.

1 Just For Laughs ... .
2 P @ P shld change to NAP
3 otherwise , they are all fully awake .
4 If true..i suggest they better dnt attend the parliament .
5 Bottom people close eye means sleeping .
6 Top people close eye and snoozing means thinking very hard .
7 ministers / MPs must take parliament session very seriously .
8 becos in parliament , very important topics are being discussed and debated .
9 must pay attention and stay awake ! !
10 sleeping on the job ?
11 His face look like wks..
12 this is becoming the PAP 's official logo
13 Sleep and Dream

Table 1: Summary sentences selected by the ILP-based method (Gillick and Favre, 2009) from a thread on criticizing parliament ministers sleeping in a meeting.

mark data sets for online forum summarization. We thus construct a data set of extractive summaries of 10 online discussion threads. Using this data set we empirically test the importance of a set of features capturing the relevance, text quality and subjectivity of candidate sentences. We find that besides relevance, two other features that are significantly important are sentence length and the number of sentiment words. We further propose some simple modifications to an ILP (Integer Linear Programming)-based summarization framework to incorporate these features and show that the modified method achieves better summarization results.

Our main contributions are the following: (1) We provide a new data set for studying extractive summarization of online discussion forums. (2) We conduct an empirical study to test the importance of several sentence features capturing text quality and subjectivity for summary sentence selection. (3) We propose modifications to a standard ILP-based extractive summarization method to incorporate good sentence features, which are shown to achieve better results.

2 Related Work

Extractive multi-document summarization: Our work is related to extractive methods for multi-document summarization, which select sentences from the original documents to form summaries. While much work has been done for this general problem, existing methods do not focus on opinion summarization. Methods for extractive multi-document summarization generally considers two factors: to increase the representativeness of the selected summary sentences with respect to the original document set, and to reduce the redundancy in the selected sentences. Existing approaches include centroid-based methods (Radev et al., 2004), learning-based methods (Kupiec et al., 1995; Wong et al., 2008) and graph-based methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Mei et al., 2010). More recently, Lin and Bilmes have done a series of work modeling text summarization with submodular functions (Lin and Bilmes, 2011; Lin and Bilmes, 2010). To globally infer an optimal set of sentences as a summary, ILP-based document summarization has been used. It was first proposed by McDonald (2007) and Gillick and Favre (2009) proposed a scalable version.

Opinion summarization: Much work on opinion summarization is for product reviews (Hu and Liu, 2004; Popescu et al., 2005; Ganesan et al., 2012). As we have pointed out, summarizing opinions from online forums, where the topics can be social issues, is quite different from summarizing product reviews. For general opinion summarization, in 2008 the Text Analysis Conference (TAC) organized an opinion summarization task. But their task is different from the one we study here. Their task is a query-oriented summarization problem where a target topic is given together with some specific questions. The corpus they use is a large set of blogs. Our task is not query-oriented, and we aim to summarize the opinions found in a single thread discussing a focused topic.

Text summarization in social media: Recently with the explosion of social media, there has been much work on summarizing social media content. In particular, much attention has been paid to Twitter summarization (Chua and Asur, 2013; Meng et al., 2012). As Twitter posts are short and not naturally organized by topics, Twitter summarization is a very different problem than ours. There has also been some studies on forum summarization (Krishnamani et al., 2013; Ren et al., 2011; Tigelaar, 2008), but the focus of these studies is not on opinion summarization.

3 Data

Since we are not aware of any existing data set that satisfies our need, we opted to create our own data. First, we picked 10 threads discussing social issues in English from the online forum Asiaone[1], which is very popular in southeast Asia. We use the first 100 posts for each thread to study our summarization problem. On average, there are 256 sentences and 3652 tokens in each thread. The vocabulary size of our data set is 5661. For each thread, we asked 3 human annotators, who are all graduate students, to carefully read the 100 posts and write a summary with a length limit of 100 words. We specifically asked the human annotators to summarize the opinions rather than facts in these threads. We also encouraged the annotators to pick sentences directly from the data but they could also compose their own sentences if necessary. In the final human summaries, there are 172 unique sentences and 156 (91%) out of them are directly picked from the original data set. We used all sentences (from all annotators) directly picked from the data set as summary sentences and all other sentences as non-summary sentences. Based on this data set, we identified discriminative features and subsequently improved our summarization method.

4 Sentence Features

In this section, we identify a number of sentence features which we hypothesize to have correlations with whether a sentence is a good summary sentence for forum opinion summarization. While a large number of features have been examined in previous studies on standard summarization (Kupiec et al., 1995; Wong et al., 2008), in this work we hypothesize that for our problem the following characteristics of a sentence are the most important: (1) representativeness with respect to the entire thread, (2) text quality, and (3) subjectivity. The first one is important for any summarization, while the last two are special for forum opinion summarization.

4.1 Representativeness

There are many different ways to measure the representativeness of a sentence with respect to the entire thread. Our objective here is not to find the best measure for representativeness but to compare the relative importance of the representativeness features with text quality features and subjectivity features for our problem. We consider two features for representativeness.

Cosine similarity. Cosine similarity has been widely used in previous summarization work (Kågebäck et al., 2014; Hu and Liu, 2004). For each sentence we calculate its cosine similarity to the entire thread, where the term vector for a sentence or for a thread is based on raw term frequency.

Concept coverage. Inspired by the concept-based ILP framework for summarization by (Gillick and Favre, 2009), we take all the bigrams (which are treated as concepts) in the original thread and use their frequencies as their weights. We then compute the weighted sum of the bigrams covered in a sentence. As the ILP-based summarization framework tries to maximize the overall concept coverage of all the selected summary sentences from one thread, a sentence with a higher concept coverage presumably is a better summary sentence candidate.

4.2 Text Quality

We hypothesize that text quality is especially important for summarizing forum posts because user-generated content tends to be of lower quality compared with traditional corpora. Typical characteristics of user-generated content that affect its text quality include use of Internet slang words, misspelling, grammatical errors, excess use of punctuation marks, etc. We hypothesize that low quality sentences are less likely to be chosen as summary sentences. While many features have been proposed to measure text quality (Pitler and Nenkova, 2008), based on our observation with our data, here we consider the following features:

4.2.1 Shallow Features

Sentence length. We use the length of a sentence in terms of the number of words as a feature. We observe that there are many short sentences in online forums and most of them do not carry much useful information. However, long sentences appear to be more informative and more useful. Thus

[1] `http://www.asiaone.com/`

we hypothesize that good summary sentences tend to be longer.

Percent of OOV (out of vocabulary) word. There exist a lot of Internet slang and abbreviations in user-generated content, such as "lol" and "hahaha." Sentences containing these words tend to be more informal and less informative. So we hypothesize that the more OOV words there are in a sentence, the less likely a sentence is a good summary sentence. Using a common English dictionary *British English Word Lists for Spell Checkers*[2], we count the number and ratio of OOV words in a sentence.

Percent of punctuation marks/emoticons. While this feature may not be important for traditional text, in user-generated content we observe that sometimes online users like to use many punctuation marks and emoticons to emphasize their emotions. We hypothesize that such sentences are not good summary sentences.

Percent of capitalized words. We also observe that in the threads we have collected, some sentences contain many capitalized words such as "HaHa" and "LOL." We hypothesize that the more capitalized words a sentence contains, the less likely it is a good summary sentence.

Average word length. This is the average length of a word in a sentence in terms of characters. With this feature, we would like to check whether good summary sentences tend to contain longer words.

4.2.2 Language Model based Feature

Log likelihood with respect to a reference corpus. Another way to measure how formal a sentence is is to use a high quality reference corpus such as a set of news articles to learn a unigram language model, and then to compute the log likelihood of generating a sentence from this language model (Pitler and Nenkova, 2008). Here we use a set of 20000 articles from Reuters as our reference corpus. Supposedly the higher the log likelihood of a sentence is, the more similar it is to the reference corpus, and we hypothesize that the more likely it is a good summary sentence. However, log likelihood is biased towards shorter sentences. We therefore take the average log likelihood per word of a sentence as our feature.

4.2.3 POS based Features

Part-of-speech based grammatical features have been widely used in text quality prediction (Feng et al., 2010; Dell'Orletta et al., 2014). They can capture the linguistic and syntactic structure of sentence, which may affect its readability. In this work, we calculate the percentage of nouns, verbs, adjectives and adverbs in each sentence.

4.2.4 Parse Tree Height

The height of the parse tree of a sentence has been used in previous work to assess text quality (Dell'Orletta et al., 2014; Pitler and Nenkova, 2008; Schwarm and Ostendorf, 2005). Here we use Stanford PCFG Parser to extract this feature. We hypothesize that as summary sentences tend to be more informative and more well-written, they may be more complicated in terms of syntactic structure and their parse tree height are probably larger than non-summary sentences.

4.3 Subjectivity

Although online forums mostly contain opinions, people sometimes also share facts or perceived facts in forums. Since our problem is opinion summarization, the summary sentences presumably should be subjective. We therefore use the following feature to test our hypothesis.

Number of sentiment words. To measure subjectivity, we take a simple approach and count the number of sentiment words in a sentence using a sentiment lexicon. We use the MPQA subjectivity lexicon (Wilson et al., 2005).

5 Feature Analysis

5.1 Approach

In Section 3 we pointed out that the sentences directly picked from the original threads by the annotators are treated as summary sentences and all other sentences are treated as non-summary sentences for the purpose of identifying useful sentence features. With the features identified in Section 4, we would like to assess the discrimination power of these features in terms of picking up summary sentences. Knowing what features are useful can help us design better summarization methods for forum opinion summarization problem. Specifically, since all our feature values are numerical, we perform unpaired Student's t-test on each feature. Student's t-test is a statistical hypothesis test, which is used to determine if two sets

[2] http://www.curlewcommunications.co.uk/wordlist.html

of data are significantly different from each other. For each feature, we get two sets of values with one set extracted from summary sentences and the other set extracted from non-summary sentences. Then we apply Student's t-test to them. If these two sets of values are significant different, the corresponding feature is useful in picking up summary sentences .

5.2 Results

Table 2 shows the results of the Student's t-test for all the features we consider. Features that show statistical significance at a 95% confidence level are marked with an asterisk.

5.2.1 Representativeness

Both features capturing representativeness of a sentence, which are cosine similarity and concept coverage, are good features. This indicates that sentences representing the salient content of a forum thread are more likely to be summary sentences. This observation follows intuition well and reflects the nature of text summarization: extracting the main content.

5.2.2 Text Quality

Shallow Features: There is much variation among the text quality features. Although we hypothesize that the features we have identified are useful, it turns out that not all of them have a statistically significant impact on whether the sentence is a good summary sentence. In particular, we find that sentence length has a positive impact. This satisfies our hypothesis that longer sentence tend to be more informative and more likely to be selected as summary sentences. The percentage of capitalized words and percentage of punctuation/emotions have negative impact. This tells us that summary sentences tend to have less capitalized words and less punctuations and emoticons. In social media, capitalized words are often used for abbreviation or emphasis and they can make a sentence less readable and less informative. Punctuations and emoticons are used more often to purely express sentiment. Sentences with higher percentage of punctuations and emoticons are less likely to contain useful information.

However, features like percent of out-of-vocabulary words and average word length can not separate summary sentences from non-summary sentences. As these two features capture the formality of words, we can see that summary sentences are similar to non-summary sentences in term of word formality. We guess that word formality is not a significant factor influencing annotators' selection of summary sentences.

Language Model based feature: The likelihood of using a language model based on Reuters corpus does not have a significant impact on selecting summary sentences. It indicates summary sentences are not more formal compared with non-summary sentences. This is consistent with the result based on shallow features.

POS based Features: In this set of features, the percent of adjectives is the only discriminative one and it has a positive impact. As our task is opinion summarization, it is intuitive that summary sentences tend to have more adjectives as many opinions are expressed by using adjectives.

Parse Tree Height: Based on the statistic test result, parse tree height is a useful feature and summary sentences tend to have larger value on this feature. This result is consistent with our hypothesis that summary sentences carry more salient content and their tree structure may appear to be more complicated.

5.2.3 Subjectivity

The simple feature of number of sentiment words in a sentence turns out to be an important feature of selecting summary sentences. This satisfies our hypothesis that summary sentences of opinions from forum should carry more opinions.

6 Forum Opinions Summarization using ILP

In the last two sections we identified and analyzed a set of sentence features to understand what characteristics good summary sentences have for our problem. While we can extend this analysis and use a supervised learning approach to classify sentences from forum posts into summary and non-summary sentences, it may not be ideal as supervised approaches suffer from their dependence on labeled training data. Moreover, even if we classify sentences into summary and non-summary sentences, we still need to consider the redundancy problem when we select sentences to form a summary. We therefore choose an unsupervised approach with a global optimization framework.

ID	feature description	p-value	test statistic
1*	cosine similarity	<0.001	6.333
2*	concept coverage	<0.001	4.695
3*	percent of punctuation/emoticons	<0.001	-4.735
4*	percent of capitalized words	<0.001	-4.190
5*	sentence length	0.001	3.438
6	average word length	0.099	1.652
7	percent of OOV	0.126	-1.530
8	average log likelihood (Reuters)	0.952	0.061
9*	percent of adjectives	0.031	2.157
10	percent of adverbs	0.176	1.353
11	percent of verbs	0.277	1.087
12	percent of nouns	0.512	-0.656
13*	parse tree height	<0.001	3.931
14*	number of sentiment words	<0.001	5.370

Table 2: Results of statistical significance tests of the features. * indicates that the result is statistically significant at a 95% confidence level. Values less than 0.001 are denoted as < 0.001.

6.1 Integer Linear Programming for Document Summarization

McDonald (2007) proposed a global optimization model to solve document summarization by integer linear programming. The idea is to maximize the overall score of selected sentences while also minimizing the redundancy among selected sentences. However, his method can have an exponentially growing number of parameters and it cannot globally measure redundancy. To handle document summarization by globally considering both content coverage and redundancy, Gillick and Favre (2009) proposed a different framework. Their objective is to cover the "concepts" in the original documents. The quality of a summary is measured by the weighted sum of concepts it covers, and a concept is counted only once regardless of how many times it occurs in the selected summary sentences. The framework therefore intrinsically handles both content coverage and redundancy reduction. The formulation is as follows:

$$\begin{aligned}\text{Maximize:}\quad & \sum_i w_i c_i \\ \text{Subject to:}\quad & \sum_j l_j s_j \le L \\ & s_j Occ_{ij} \le c_i, \forall i, j \\ & \sum_j s_j Occ_{ij} \ge c_i, \forall i \\ & c_i \in \{0,1\} \quad \forall i \\ & s_j \in \{0,1\} \quad \forall j\end{aligned}$$

where w_i is the weight of concept i, c_i is a binary indicator for concept i which will be set to 1 when i is covered by the summary. s_j is the binary indicator for sentence j which is 1 when the sentence is selected as a summary sentence. Occ_{ij} is a binary variable indicating the occurrence of concept i in sentence j, which would be 1 if i occurs in j. l_j is the length of sentence j. We need to solve optimization problem and get the optimal values of c_i and s_j for all i and j.

6.2 Our Modifications

We can see that the above framework does not consider sentence quality or subjectivity. Based on the findings from Section 5, we propose the following modifications to the concept-based ILP framework.

LengthMod-1: Since we find that summary sentences from forums tend to be longer, we propose to minimize the total number of sentences in the summary as follows:

$$\text{Maximize:} \sum_i w_i c_i - \lambda \sum_j s_j.$$

where λ is a free parameter in all three modifications. The second term is essentially the total number of sentences selected. The other constraints for the optimization problem remain the same.

LengthMod-2: Alternatively, we propose the following objective function to favor longer sentences:

$$\text{Maximize:} \sum_i w_i c_i + \lambda \sum_j l_j^2 s_j.$$

With the total length of all selected sentences capped at L, the second term above favor the selection of fewer, longer sentences.

SubjectMod-1: To favor sentences with subjective words, we can formulate the following objective function:

$$\text{Maximize: } \sum_i w_i c_i + \lambda \sum_j o_j s_j,$$

where o_j is the sentiment score for sentence j, which is computed by counting the number sentiment words in j.

SubjectMod-2: Alternatively, to model the influence of sentiment words, we can change the way w_i is calculated. In the study by Gillick and Favre (2009), w_i is the frequency of concept i in the input documents. Here, we change it to be the frequency of i appearing in opinionated sentences. For simplicity, we treat sentences containing sentiment words as opinionated sentences. The intuition of the original method is try to cover as many frequent concepts as possible. The intuition of ours is to cover as many *opinion related* concepts as possible.

6.3 Results

	ROUGE-1	ROUGE-2
Baseline	0.3418	0.1062
LengthMod-1	0.3483	0.1187
LengthMod-2	0.3469	0.1182
SubjectMod-1	0.3399	0.0991
SubjectMod-2	**0.3576**	**0.1191**

Table 3: Summarization Performance

To test the effectiveness of our modifications, we applied both them and the baseline method on the forum data introduced in Section 3. The human editted summaries are used as the gold standard references. For our modifications, when summarizing one thread, we use all other 9 threads and the corresponding human summaries as training data to find the optimal λ. We use ROUGE-1 and ROUGE-2 as the evaluation metric.

In Table 3 we show the performance of the baseline method and our modifications. We can see that modifications that incorporate length into the objective function both give better performance over the baseline. This shows that our modified versions of the objective function can effectively bring in longer sentences for summaries. However, the two modified methods based on sentence subjectivity have very different performance. While SubjectMod-2 outperforms the baseline (and all other modifications), SubjectMod-1 does not outperform the baseline.

A deeper analysis of SubjectMod-1 and SubjectMod-2 can reveal their difference. SubjectMod-2 changes the way concept weights are calculated. In this method, concepts co-occurring more with sentiment words in the same sentence will be more important. The algorithm tries to cover as many sentiment related frequent concepts as possible. Coverage and subjectivity are incorporated and considered at the same time. However, SubjectMod-1 considers coverage and subjectivity separately. If a sentence contains some frequent but not opinion related concepts and a few sentiment words, it may be selected as a summary sentence by SubjectMod-1.

7 Conclusions

In this paper, we studied the problem of summarizing opinions from online forum threads. We first constructed a data set with human generated model summaries and then identified a number of sentence features which we hypothesized to be useful in characterizing good summary sentences. These features cover representativeness, text quality and subjectivity of a sentence. Based on the model summaries we have obtained, we evaluated the effectiveness of these features based on Student's t-test. We found that a number of these features are significantly discriminative in identifying summary sentences. We then proposed to modify an ILP-based summarization framework to take sentence length and subjectivity into consideration.

Our study provides insight into the general problem of summarizing online opinions from forum discussions, which has not been well studied. Our findings suggest that a number of factors other than content coverage are important to consider when it comes to summarizing opinions from social media. Our proposed modifications to a principled summarization framework show promising results. Our study is still preliminary. In the future, we plan to study how to further improve the ILP-based summarization framework to incorporate more considerations. We also expect that 1) it is useful to use fine-grained opinion extraction to extract and normalize opinions before they can be summarized, 2) social media properties like users' attributes and social effect can be helpful in summarizing text content.

References

Freddy Chong Tat Chua and Sitaram Asur. 2013. Automatic summarization of events from social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media*, pages 81 – 90.

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. 2014. Assessing the readability of sentences: Which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, page 163C173.

Günes Erkan and Dragomir R. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365–371.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284.

Kavita Ganesan, ChengXiang Zhai, and Evelyne Viegas. 2012. Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web*, pages 869–878.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*, pages 31–39.

Janani Krishnamani, Yanjun Zhao, and Rajshekar Sunderraman. 2013. Forum summarization using topic models and content-metadata sensitive clustering. In *Web Intelligence/IAT Workshops*, pages 195–198.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 510–520.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European Conference on Information Retrieval Research*, pages 557–564.

Q. Mei, J. Guo, and D. Radev. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. 2012. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–387.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

Ana-Maria Popescu, Bao Nguyen, and Oren Etzioni. 2005. Opine: Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346.

D.R. Radev, H. Jing, M. Styś, and D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Zhaochun Ren, Jun Ma, Shuaiqiang Wang, and Yang Liu. 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*, pages 879–884.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530.

Almer S. Tigelaar. 2008. *Automatic discussion summarization : a study of Internet forums*. Ph.D. thesis, University of Twente.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22Nd International Conference on Computational Linguistics*, pages 985–992.