Summarizing Newspaper Comments

Clare Llewellyn, Claire Grover and Jon Oberlander
The School of Informatics, University of Edinburgh
Edinburgh
United Kingdom

Abstract

This work investigates summarizing the conversations that occur in the comments section of the UK newspaper the Guardian. In the comment summarization task comments are clustered and ranked within the cluster. The top comments from each cluster are used to give an overview of that cluster. It was found that topic model clustering gave the most agreement when evaluated against a human gold standard. This approach is compared to cosine distance clustering and k-means clustering. PageRank was found to be the prefered ranking system when compared with TF-IDF, Mutual Information gain and Maximal Marginal Relevance and evaluated against sets of comments summarized by a journalist for the Guardian letters page.

Introduction

This work investigates summarizing the conversations that occur in the comments section of the UK newspaper, the Guardian.

This comment system, like many others, allows users to view comments either in a temporal fashion, oldest or newest first, or as threads. The thread style of presentation means that often users only look at recent or popular content - this can give a misleading impression of the overall discussion. The context and sense of variety of the discussion may be lost, leading to users repeating previous discussions or not adding their comments to the appropriate thread. This type of structure may not be the best way to initially interact with this type of data.

Newspapers can accumulate many hundreds and sometimes thousands of comments. On initial viewing these comments may seem less than useful, full of replication, extreme views, petty arguments and spam, but when studied closely and analyzed effectively they provide multiple view points and a wide range of experience and knowledge from many different sources. If we can find ways to analyze the information correctly we can exploit this crowd-sourced information aggregation. Summarizing the content of these comments allows users to interact with the data at a higher level. It gives an overview impression of the conversation that has occurred.

A survey paper by (Potthast et al. 2012) suggests that the most important tasks with regard to understanding the information available in comments are filtering, ranking and summarizing the comments. This work aims to explore and compare current approaches to the summarization of comments.

The summarization domain is well developed. The earliest focus of the field was single document summarization (Gupta and Lehal 2010), this approach was extended to summarization of multiple documents on the same topic (Goldstein et al. 2000) and to summarizing discussions such as email conversations (Cselle, Albrecht, and Wattenhofer 2007). Within the social media domain some examples of current summarization work use blogs (Hu, Sun, and Lim 2007), tweets (Chakrabarti and Punera 2011) and reviews (Brody and Elhadad 2010).

The basic idea behind summarization of textual data is the grouping together of similar information and describing those groups (Rambow et al. 2004). Once these groups are formed they are described using either an extractive or abstractive approach. Extractive summarization uses units of text, generally sentences, from within the data in the group to represent the group. Abstractive summarization creates a description of the data in the group as a whole analogous to the approach a human would take if they were doing the task. Abstractive summarization is a very complex task. As comment summarization is a task, which has only recently been attempted, there is a focus on extractive approaches.

In this paper we focus on extractive summarization of multiple documents, grouped by topic. In similar work latent Dirichlet allocation (LDA) topic modeling (Blei, Ng, and Jordan 2003) has been used as a basis for grouping topics. We investigate this approach by comparing topic model clustering with other clustering approaches (clustering by key words, cosine distance and k-means) and we assess these approaches against a gold standard produced by humans. Once grouped, we extract part of the content to describe the groups by ranking the comments within their clusters. Several ranking methods are compared (TFI-DF, Mutual Information Gain, Text-Rank, and Maximal Marginal Relevance). These approaches are then evaluated by human participants who compare the summarization sets against comments summarized by a journalist for the Guardian letters page.

Previous Work

The aim of this work is to compare previous approaches with each other on new content and perform an evaluation against a human gold standard. Comment summarization is a new domain and therefore there is a limited amount of previous work this area. We focus on two papers, Ma et al. (2012) and Khabiri, Caverlee, and Hsu (2011). Ma et al. summarize discussion on news articles from Yahoo! News and Khabiri, Caverlee, and Hsu summarize comments on YouTube videos. They both agree on the definition of the basic task as clustering comments into topics, ranking to identifying comments that are key in the clusters and evaluating the results through a human study.

Clustering Both approaches focus on using LDA topic modeling to cluster the data. Ma et al. investigates using two topic models; one where topics are derived from the original news article and an extended version that allows new topics to be formed from the comments. They found that the extended version was judged superior by a user study. Khabiri, Caverlee, and Hsu contrasted LDA topic models with k-means and found topic modeling superior.

Ranking Khabiri, Caverlee, and Hsu investigated two approaches to ranking comments, scoring important terms by Term Frequency Inverse Document Frequency (TF-IDF) and Mutual Information (MI) and a PageRank style random walk across a graph built on similarity-based relationships amongst sentences. They found their performance of PageRank superior to the other approaches. Ma et al. ranked by Maximal Marginal Relevance (MMR) and Rating and Length (RL) (user rating * length). They found that MMR gave the better performance.

Evaluation Generally summaries are evaluated against a ground truth summary generated by humans. Such ground truth summaries are not widely available in the comment summarization domain and, due to the size of the comment sets, a human version of this task would take a not insignificant amount of time. Therefore other evaluation techniques have been investigated. Ma et al. and Khabiri, Caverlee, and Hsu both conducted user studies to evaluate their summaries.

The Ma et al. study asked three users to evaluate topic cohesion, topic diversity, and news relatedness by rating comments from 1 (lowest) to 5 (highest) for each feature for each summary. The summaries contained 15 comments that were generated from the top 5 ranked comments from the top 3 largest clusters. Khabiri, Caverlee, and Hsu asked five users to give a score of 0-5 to interesting and informative comments in a set made up of the first 50 comments taken from 30 videos. So each comment would have a score of 0-5 depending on how informative and interesting the users thought the comment. This was used to create a gold standard ranking for the comments on each video to compare with the automated approach.

Experiments

Data

Our corpus is from the Guardian newspaper. It is composed of online comments created by readers. These readers have to register and post under a username. The site is lightly modurated and responds to users' complaints. The comments sections are open for a variable amount of time after the article is published and are then closed. The comments are harvested from the site after the comment section is closed.

Clustering

There is general agreement that some variation of topic modeling is the best approach for clustering comments into topics. In order to confirm this claim we compared topic modeling to several other clustering techniques.

For the clustering experiments we used the comments responding to an article reviewing the iPad mini (http://gu.com/p/3bb88). There were 161 comments produced over 2 days. To produce the gold standard data two humans were asked to assign comments into groups of the same topic. No guidance was given as to the number of topics required. Annotator A determined that there were 26 topics whereas annotator B identified 45 topics. While this seems a large difference, it was entirely due to a variation in numbers of singleton clusters. With the singleton clusters removed both users had created 14 clusters.

The results that are given are in terms of a micro averaged F-Score. The F-score is the harmonic mean of precision and recall. To use this measure with data that has been assigned to multiple clusters a micro-average F-score is used. This gives an average F-score across comments. As this gives equal weight to every comment it favors larger clusters. In this task getting the bigger clusters right, those with more information is most important. The macro average score (an average across clusters) is not shown here and is substantially lower in each case. For more details on this metric see (Sokolova and Lapalme 2009).

The human-human micro averaged F-Score was 0.6066 including the singletons and 0.805 without. This data was taken as the gold standard against which the other clustering techniques were evaluated. The clustering techniques used were clustering by unigrams, cosine distance and k-means.

Not all approaches worked best with 14 clusters. In table 1 the results shown are the best reguardless of cluster number. Details are stated when the number of clusters is not 14.

Baseline the unigram approach is used as a baseline. A list of the 14 most frequent terms is extracted from the set. The comments are assigned to the group of the most frequent term from the list within that comment.

Cosine distance clustering creates a vector representation of the text in the comment. The comments are clustered into groups of comments that have a low cosine distance. This approach is popular in particular with social media data (Becker, Naaman, and Gravano 2010) as it allows clustering in a single pass whereas most clustering techniques require multiple iterations in order to determine the best clusters. This approach is implemented using code from the COSMOS project (http://www.cosmosproject.net/). For the results shown in table 1 the threshold value for group inclusion is 0.1 and the number of features is set at 20. This gives 9 clusters.

K-means clustering is a more traditional approach. Again it requires the creation of vectors to represent the comments. Random points are selected within the vector space and comments allocated to the closest points (in this case using a Euclidean distance measure). The points are then moved to the centre vector of the comments in that cluster and the comments are all reassigned to the nearest point. This process in this case is repeated 20 times and implemented using k-means clustering from the Natural Language Tool Kit (Bird, Klein, and Loper 2009). For the results shown in the table 1 a value of 20 is given for k with 4 singletons removed leaving 16 clusters.

LDA topic modeling is a generative model produced to determine the topics contained in a text document. A topic is formed from words that often co-occur. Therefore the words that occur more frequently across multiple documents are most likely in the same topic. It is also true that each document may contain a variety of topics. LDA provides a score for each document for each topic. In this case we assign the document to the topic for which it has the highest score. This approach was implemented using the Mallet tool kit (Mccallum 2002) the parameters used were set with 14 topics (clusters), an optimization interval of 10 and an initial alpha 5 and beta 0.01.

Clustering Results As seen in Table 1 the topic modeling was found to provide the most human like clusters. No other approach beat the unigram baseline approach.

Approach	Human 1	Human 2	Average
Human-Human	0.805		
Unigrams	0.330	0.313	0.321
K-means	0.200	0.209	0.204
Cosine Distance	0.202	0.201	0.201
Topic Models	0.431	0.461	0.446

Table 1: Micro-Averaged F-score for Clustering Methods

Ranking

The aim of this experiment was to compare various methods for ranking the comments in the clusters.

The comments used for this experiment were taken from an article discussing a former London gang leader (http://gu.com/p/3yv6n). There were 136 comments which were clustered using topic modeling. The resulting clusters were ranked using the following metrics:

TF-IDF is a widely-used metric for determining important terms in text. In this case this measure indicates how much information each term contributes to the cluster. For each term the TF part of the metric is the number of time the term appears in the comments of a specific cluster normalized by the total number of terms in that cluster. The IDF is the logarithm of total number of comments divided by the number of comments that the term appears in. An average TF-IDF score is computed for each comment in the cluster by averaging the score of the terms in that comment. As this approach tends to favor short comments with a few very important terms a second approach is also used that penalizes the shorter comments (refered to as TF-IDF long).

Mutual Information Gain (MI) is similar to the TF-IDF metric as it measures the amount of information each term provides to the cluster. (Hsu, Khabiri and Caverlee (2009)). The MI is computed for each comment using the following equation taken from the Khabiri, Caverlee, and Hsu:

$$MI = \frac{N_{11}}{N} log_2 \frac{N_{11}N}{N_{1.}N_{.1}} + \frac{N_{10}}{N} log_2 \frac{N_{10}N}{N_{1.}N_{.0}} +$$

$$\frac{N_{01}}{N} log_2 \frac{N_{01}N}{N_{0.}N_{.1}} + \frac{N_{00}}{N} log_2 \frac{N_{00}N}{N_{0.}N_{.0}}$$

N is the number of terms, the first subscript tells us if the comment contains the term (1) or not (0) where as the second subscript indicates if it is this cluster (1) or all other clusters(0). The MI for the comment is then determined by dividing the overall score by the number of terms.

PageRank many summarization approaches recognize repetition as an important factor when choosing the sentences to use in an extraction approach. PageRank is a graph-based ranking process and involves a random walk over a graph produced by creating links between comments that repeat earlier terms from previous comments. A comment that repeats text from an earlier comment gives an indication that this is a supported comment within the cluster and thereby indicates more important comments. The walk determines scores for each comment by propagating scores through the network.

To implement this metric we use the Python package NetworkX (http://networkx.github.io/). Each comment is compared to other comments in the cluster to see if they share more than 10 terms. If so a directed link is created from the comment to the earlier comment in the graph. The random walk is performed until convergence to find comments with more support from the later comments. This approach favours longer comments with more terms therefore a second approach is used which weights links according to a fraction of the number of terms in both comments (refered to as PageRank short).

Maximal Marginal Relevance another approach to find the best comments for extraction is by identifing the comments closest to the centroids of the clusters.

$MMR = \lambda(Sim_1)(D_i, C) - (1 - \lambda)Sim_2(D_i, D_j)$

Maximal Marginal Relevance (MMR) does this by comparing a vector representation of the centroid with those created from the comments using a similarity measure (Sim in the equation), we use cosine similarity for both similarity measures. D_i is the centroid of the cluster and D_j the centroid of the summary, and C the vector of the comment. We used a λ of 0.9. Once a comment is added to the summary (ranked as closest to the centroid) a penalty is imposed on further comments weighted on their similarity to comments already in the summary, thereby increasing diversity. For futher information see Goldstein et al. (2000) or Ma et al. (2012).

Evaluation of Ranking

The print version of the Guardian newspaper produced on a Saturday has a section on the letters page containing a human-produced summary from the set of comments discussing a selected article. We use this as a gold standard human-produced summary. Each week the human produced

data is clustered into 3 or 4 topics. Unfortunately the comments in each topic are not available only the final summary. The human-produced summary is shown below:

```
40% criticised gang culture for creating
a desire for fame and respect
33% would like to hear more from victims
of gang violence
17% found Dagrou's story depressing
10% believed he should be praised for
turning his life around
```

Once the data had been clustered using topic modeling each metric described above was used to select three comments from each of the clusters to represent that cluster. The text unit extracted in this case, conforming to the approach taken in similar work, is a comment rather than a sentence. As a baseline three random comments were selected from each cluster.

Six participants were asked to read the gold standard and then asked to compare it with the automatically generated summaries. They were asked to rank the summaries from 1 to 7 with 7 being the best and 1 the worst. In addition they were asked to comment on the summaries.

Ranking Results

As we can see in Table 2 PageRank and PageRank short are judged to be the best performing ranking mechanisms closely followed by MMR and MI. The TF-IDF metric is judged as worse than random in for TF-IDF and the same as random for TF-IDF long.

Ranking Method	Number of participants ranking in this class %							Ave. Rank
	7	6	5	4	3	2	1	
TF-IDF	0	0	0	0	0	17	83	1
TD-IDF(long)	0	0	17	0	33	33	17	2.29
MI	0	0	33	50	17	0	0	3.57
PageRank	50	33	17	0	17	0	0	5.86
PageRank(short)	33	33	17	0	17	0	0	4.86
MMR	17	17	33	33	0	0	0	4.43
Random	0	0	0	17	33	50	0	2.29

Table 2:Ranking of Ranking Methods (higher is better)

In general the participants commented that they did not like summaries made up of very long or very short comments. They also commented that the summaries in general were quite poor.

Conclusion and Future Work

The contributions of this work are confirming that LDA-topic modelling is the best approach for clustering comment data into topics and that the prefered ranking mechanism, with this this limited number of comments, was found to be PageRank.

We also found that it was also clear that the participants did not think that the summaries were good when compared to human-produced summaries. It would be useful to take advantage of advances from the wider summarization field, in particular those from product review and email summarization to improve this. In particular methods that may help include building more accurate initial topic clusters, extracting sentences or phrases rather than full comments, and using sentiment analysis to indicate polarity.

References

Becker, H.; Naaman, M.; and Gravano, L. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, 291–300. ACM.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python*. O'Reilly Media.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Brody, S., and Elhadad, N. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies*, 804–812. Association for Computational Linguistics.

Chakrabarti, D., and Punera, K. 2011. Event summarization using tweets. In *ICWSM*.

Cselle, G.; Albrecht, K.; and Wattenhofer, R. 2007. Buzztrack: topic detection and tracking in email. In *Proceedings of the 12th international conference on Intelligent user interfaces*, 190–197. ACM.

Goldstein, J.; Mittal, V.; Carbonell, J.; and Kantrowitz, M. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization-Volume 4*, 40–48. Association for Computational Linguistics.

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

Hu, M.; Sun, A.; and Lim, E.-P. 2007. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 901–904. ACM.

Khabiri, E.; Caverlee, J.; and Hsu, C.-F. 2011. Summarizing user-contributed comments. *ICWSM*.

Ma, Z.; Sun, A.; Yuan, Q.; and Cong, G. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 265–274. ACM.

Mccallum, A. 2002. MALLET: a machine learning for language toolkit.

Potthast, M.; Stein, B.; Loose, F.; and Becker, S. 2012. Information retrieval in the commentsphere. *ACM TIST* 3(4):68.

Rambow, O.; Shrestha, L.; Chen, J.; and Lauridsen, C. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, 105–108. Association for Computational Linguistics.

Sokolova, M., and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45(4):427–437.