Improving Topic Model Clustering of Newspaper Comments for Summarisation

Won't somebody please think of the children? Improving Topic Model Clustering of Newspaper Comments for Summarisation

Clare Llewellyn
School of Informatics
University of Edinburgh
Edinburgh, UK
s1053147@sms.ed.ac.uk

Claire Grover
School of Informatics
University of Edinburgh
Edinburgh, UK
grover@inf.ed.ac.uk

Jon Oberlander
School of Informatics
University of Edinburgh
Edinburgh, UK
jon@inf.ed.ac.uk

Abstract

Online newspaper articles can accumulate comments at volumes that prevent close reading. Summarisation of the comments allows interaction at a higher level and can lead to an understanding of the overall discussion. Comment summarisation requires topic clustering, comment ranking and extraction. Clustering must be robust as the subsequent extraction relies on a good set of clusters. Comment data, as with many social media datasets, contains very short documents and the number of words in the documents is a limiting factors on the performance of LDA clustering. We evaluate whether we can combine comments to form larger documents to improve the quality of clusters. We find that combining comments with comments that reply to them produce the highest quality clusters.

1 Introduction

Newspaper articles can accumulate many hundreds and sometimes thousands of online comments. When studied closely and analysed effectively they provide multiple points of view and a wide range of experience and knowledge from diverse sources. However, the number of comments produced per article can prohibit close reading. Summarising the content of these comments allows users to interact with the data at a higher level, providing a transparency to the underlying data (Greene and Cross, 2015).

The current state of the art within the comment summarisation field is to cluster comments using Latent Dirichlet Allocation (LDA) topic modelling (Khabiri et al., 2011; Ma et al., 2012; Llewellyn et al., 2014). The comments within each topic cluster are ranked and comments are typically extracted to construct a summary of the cluster. In this paper we focus on the clustering subtask. It is important that the clustering is appropriate and robust as the subsequent extraction relies on a good set of clusters. Research in a related domain has found that topical mistakes were the largest source of error in summarising blogs – (Mithun and Kosseim, 2009) a similar data type.

Comment data, as with many social media datasets, differs from other content types as each 'document' is very short. Previous studies have indicated that the number of documents and the number of words in the documents are limiting factors on the performance of topic modelling (Tang et al., 2014). Topic models built using longer documents and using more documents are more accurate. Short documents can be enriched with external data. In our corpus the number of comments on each newspaper article is finite and the topics discussed within each set have evolved from the original article. We therefore decided not to increase the set with data from external sources.

In this work we consider whether we can combine comments within a comments dataset to form larger documents to improve the quality of clusters. Combining comments into larger documents reduces the total number of comments available to cluster which may decrease the quality of the clusters. The contribution of this work is in showing that combining comments with their direct replies, their children, increases the quality of the clustering. This approach can be applied to any other task which requires clustering of newspaper comments and any other data which contains small documents linked using a thread like structure. Combining data in this way to improve the clustering reduces the need to import data from external sources or to adapt the underlying clustering algorithm.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 43–50,
Berlin, Germany, August 7-12, 2016.

2 Related Work

2.1 Summarisation

The summarisation domain is well developed. The earliest focus of the field was single document summarisation – for a survey paper see (Gupta and Lehal, 2010). This approach was extended into the summarisation of multiple documents on the same topic (Goldstein et al., 2000) and to summarising discussions such as email or Twitter conversations (Cselle et al., 2007; Sharifi et al., 2010; Inouye and Kalita, 2011).

The basic idea behind the summarisation 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 summarisation uses units of text, generally sentences, from within the data in the group to represent the group. Abstractive summarisation creates a description of the data in the group as a whole, analogous to the approach a human would take.

2.1.1 Comment Summarisation

Abstractive summarisation is a very complex task, and because comment summarisation is a relatively new task, current work mostly focuses on extractive approaches. The general task involves clustering the comments into appropriate topics and then extracting comments, or parts of comments to represent those topics (Khabiri et al., 2011; Ma et al., 2012). Ma et al. (2012) summarise discussion on news articles from Yahoo!News and Khabiri et al (2011) summarise comments on YouTube videos. Both studies agree on the definition of the basic task as: clustering comments into topics, ranking to identify comments that are key in the clusters, and evaluating the results through a human study. Both approaches focus on using LDA topic modelling to cluster the data. Ma et al. (2012) explored two topic models, one where topics are derived from the original news article and a second, extended version that allows new topics to be formed from the comments. They found that the extended version was judged superior in a user study. Khabiri et al. (2011) contrasted LDA topic models with k-means and found topic modelling superior. A study by Llewellyn et al. (2014) contrasted topic modelling, k-means, incremental one pass clustering and clustering on common unigrams and bigrams. They found that the topic modelling approach was superior. Aker et al. (2016) looked at a graph based model that included information from DBpedia, finding that this approach out performed an un-optimised LDA model. They then labelled the clusters using LDA clustering and extracted keywords.

Other work has been conducted in related domains such as summarising blogs, microblogs and e-mail.

2.1.2 Blog Summarisation

Comments are similar to blogs in that they are generated by multiple individuals who exhibit a vast array of writing styles. Mithum and Koseim (2009) found that whereas news articles have a generalisable structure that can be used to aid summarisation, blogs are more variable. In particular they found that errors in blog summarisation are much higher than in news text summarisation. They determined that errors were often due to the candidate summary sentences being off topic and they suggest that blog summarisation needs to be improved in terms of topic detection. When investigating the summarisation of blogs and comments on blogs Balahur et al.(2009) found that it is very common to change topics between the original blog post and the comments, and from comment to comment. The research of Mithum and Koseim (2009) and Balahur et al. (2009) indicates that topic identification is a key area on which to concentrate efforts in the emerging field of comment summarisation.

2.1.3 Microblog Summarisation

A significant amount of work has been conducted in the area of Twitter summarisation. Many Twitter summarisation techniques exploit that tweets often include hashtags which serve as an indication of their topic. Duan et al.(2012) designed a summarisation framework for Twitter by defining topics and selecting tweets to represent those topics. The topics are defined using hashtags and are split when at high volume by specific time slices and word frequency. Rosa et al. (2011) also use hashtags to cluster tweets into topics, using them as annotated classes for training data. They focus on supervised machine learning, specifically SVM and K Nearest Neighbour, as they found the results from unsupervised clustering (LDA and k-means clustering) performed poorly when applied to Twitter data. In a further Twitter summarisation

tool, TweetMotif, O'Connor et al. (2010) use language modelling to create summaries. They form topic clusters by identifying phrases that could define a topic, looking for those phrases in the corpus and merging sets of topics that are similar. Research on microblog summarisation indicates that when summarising comments it is possible but difficult to use unsupervised clustering and several rules have been suggested that can be followed to produce the most suitable clusters for summarisation.

2.1.4 E-mail Summarisation

E-mail and comments are similar in several respects: they both exhibit a thread-like structure, containing multiple participant conversations that occur along a variable time line, they may refer back to previous parts of the conversation and exhibit high variability in writing styles (Carenini et al., 2007). Topic identification is challenging in e-mail threads. Wan and Mckeown (2004) noted that several different tasks were conducted in email conversations: decision making, requests for action, information seeking and social interaction. Rambow et al. (2004) found that e-mail has an inherent structure and that this structure can be use to extract e-mail specific features for summarisation. This suggests that comments may have an inherent structure which can be used to assist in summarisation.

3 Methods

3.1 Data

The work reported here is based on comments from news articles taken from the online, UK version of the Guardian newspaper. It is composed of online comments that are created by readers who have registered and posted under a user-name. The site is moderated and comments can be removed. We harvested the comments once the comment section is closed and the data is no longer updated. The comment system allows users to view comments either in a temporal fashion, oldest or newest first, or as threads. Users are then able to add their own comments to the set by either posting directly or by replying to another user, adding their comments to any point in the thread. This develops a conversational style of interaction where users interact with each other and comment upon the comments of others. The topics discussed can therefore evolve from the topics of the original article.

In total we have gathered comments posted in response to thirteen articles. Each week a journalist from the Guardian summarises the comments on one particular article and we have selected data from these weekly summaries to provide a further point of comparison. A typical example is our comment set 5 where the initial article was titled 'Cutting edge: the life of a former London gang leader', the journalist had divided the comments into sets as follows:

- 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

An example of a comment that fit into the journalist based classification scheme is: *"I'd love to see an in-depth article about a person whose life is made a complete misery by gangs. You know, maybe a middle-aged lady who lives on her own in a gang area, something like that."*

An example of a comment that does not fit into the classification scheme: *"So people who don't have to turn their lives around are ignored and not supported. These are the people who are sometimes homeless cause there is no help if you haven't been in prison or don't have kids"*.

In this work we refer to all of the comments on a single article as a *comment set*. There is data that has been annotated by humans (the gold standard set) and data that has not. The gold standard data set contained three comment sets. It was produced by human(s) assigning all comments from a comment set to topic groups. For one comment set two humans assigned groups (Set 1) and for two comment sets (Sets 2 and 3) a single human assigned groups. No guidance was given as to the number of topics required, but the annotators were asked to make the topics as broad or as inclusive as they could.

In the set where both humans assigned topics the first annotator determined that there were 26 topics whereas the second annotator identified 45 topics. This difference in topic number was due to a variation in numbers of clusters with a single

Table 1: Comment Set Composition - A description of the data set

Set	1	2	3	4	5	6	7	8	9	10	11	12	13
Comments	160	230	181	51	121	169	176	205	254	328	373	397	661
Authors	67	140	112	28	65	105	103	111	120	204	240	246	420
Threads	54	100	82	21	53	71	67	80	95	132	198	164	319
Groups of siblings	126	186	154	45	108	139	148	160	205	256	314	320	553
Time Segment	77	113	68	33	72	110	76	117	142	160	124	119	203
Over 50 Words (%)	58	52	29	39	37	36	18	38	49	44	26	26	30
Mean number of words	80	81	45	58	53	45	38	69	72	61	40	43	48
Human topics	14	21	20	-	-	-	-	-	-	-	-	-	
Automatic topics	-	-	-	5	5	7	8	5	5	18	18	16	7

member. Once these were removed both annotators had created 14 clusters. The human-human F-Score was 0.607 including the single clusters and 0.805 without. It was felt that agreement at this level meant that double annotation was not required for the futher two sets. All annotated sets have the clusters with single members removed.

A further 10 sets of comments were collected which were not annotated. Table I shows the composition of these comment sets. We can see that the number of comments varies and that number of authors, threads, groups of siblings (comments that reply to the same comment) and time segments tend to increase with size. The number of words in a comment does not. The sets of comments, with annotations where available, can be found at (Llewellyn, 2016).

3.2 Data Manipulation

We have investigated methods for combining the individual comments into larger 'documents' using metadata features. The data is combined according to aspects extracted from the metadata; these are as follows:

- STANDARD: Comments are not combined in any way. This is a baseline result to which the other results can be compared.
- AUTHOR: Comments are grouped together if they have the same author. A common approach to increase the size of Twitter documents is to group tweets together that come from a single author on the assumption that authors stick to the same/similar topics. Here the same approach is tried with comments.
- TIME: Comments are grouped together within a ten minute segment. Comments may be on the same topics if they are posted at the same time (if the users are viewing comments through the newest first method).

It is hypothesised that there may be topical consistency within threads. The 'threadness' was identified in several ways:

- FULL THREAD: Comments were grouped together to reflect the full thread from the original root post and including all replies to that post and all subsequent posts in the thread.
- CHILDREN: A comment is grouped with all direct replies to that comment.
- SIBLINGS: A comment is grouped with its siblings, all other comments that reply to a specific comment.

All of the groups of related comments are combined together, according to the method, to form a single document for each group.

3.2.1 Short Documents

Previous work indicates that removing short documents from the data sets prior to topic modelling improves the quality of the topic models (Tang et al., 2014). We found, in an experiment into whether length of comments influenced the quality of clusters, that removing comments that contain few than 50 terms increases the ability of a topic model to classify documents that are longer than 50 terms but it does not increase the ability to classify all documents, especially shorter documents. If we deem it useful to have short comments in the clusters for the ranking and extraction phase of summarisation, then it is important that these shorter documents are retained in the model

building stage, we therefore include them in our experiments detailed here.

3.3 Topic Modelling

The clustering method used in this work is Latent Dirichlet Allocation (LDA) topic modelling (Blei et al., 2003). It produces a generative model used to determine the topics contained in a text document. A topic is formed from words that often co-occur, therefore the words that co-occur more frequently across multiple documents most likely belong to 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 or topics for which it has the highest score.

This approach was implemented using the Mallet tool-kit (McCallum, 2002). The Mallet tool kit topic modelling implementation allows dirichlet hyper-parameter re-estimation. This means that although the hyper parameters are initially set it is possible to allow them to be re-estimated to better suit the data set being modelled. In these experiments, after previous optimisation tests, we initially set the sum of alpha across all topics as 4, beta as 0.08. We set the number of iterations at 1000, and we allow re-estimation of the dirichlet hyper-parameters every 10 iterations.

In order to cluster the comment data into topics an appropriate number of topics must be chosen. In choosing the number of topics we aim to pick a number which strikes a balance between producing a small number of broad topics or a large number of overly specific topics. We aim to echo a human like decision as to when something is on- or off-topic. Too few items in each topic is to be avoided, as is having a small number of topics (O'Connor et al., 2010).

In our data set, we choose the number of clusters by two methods. When data has been annotated by humans the number of topics identified by humans was chosen as the cluster number. When the data had not been annotated by humans the cluster number was identified using an automatic method of stability analysis. This method was proposed by Greene, O' Callaghan, and Cunningham (2014), and it assumes that if there is a 'natural' number of topics within a data set, then this number of topics will give the most stable result each time the data is re-clustered. Stability is calculated using a ranked list of most predictive topic words. Each time the data is modelled, the change in the members and ordering of that list is used to calculate a stability score. Green et. al (2014) used the top twenty features to form their ranked list of features. Here, as the length of the documents is shorter, we use the top ten.

The sets of documents as described in the previous sections are then used to build topic models and the comments are assigned to topical clusters using these models. Ten-fold cross-validation is used. As topic modelling is a generative process, the topics produced are not identical on each new run as discussed in more detail in (Koltcov et al., 2014). Therefore the process is repeated 100 times, so that an average score can be supplied.

Table 2: Combined Data, Annotated, F-score (results that beat the standard baseline are in bold)

	1	2	3
Standard Baseline	0.59	0.36	0.33
Author	0.43	0.34	0.32
Children	**0.70**	**0.41**	**0.48**
Full Thread	**0.63**	**0.38**	**0.37**
Siblings	0.59	0.37	0.33
Time	0.38	0.31	0.24

3.4 Metrics

There are two main metrics that are exploited in this work: Perplexity and micro-averaged F-score. Perplexity is judged by building a model using training data, and then testing with a held out set to see how well the word counts of the test documents are represented by the word distributions represented in the topics in the model (Wallach et al., 2009). This score shows how perplexed the model is by the new data. Perplexity has been found to be consistent with other measures of cluster quality such as point-wise mutual information (Tang et al., 2014). PMI data is also available and can be supplied if requested.

It is difficult to judge when a perplexity score is 'good enough' as perplexity will continue to decrease as the number of clusters increases. Topic models that represent single comments are the least perplexed. Therefore a section of the dataset has been hand annotated and this is used to provide a micro-averaged F-score. This can be used to gauge if the perplexity scores are equivalent to human judgements. For more details on this metric see Sokolova and Lapalme (2009).

Table 3: Combined Data - Perplexity Score (the best / least perplexed model is in bold)

Comment Set	1	2	3	4	5	6	7	8	9	10	11	12	13
Standard	**253**	671	520	343	555	444	531	960	1084	818	**659**	756	810
Author	572	644	525	422	555	582	518	1005	1224	908	669	766	761
Children	373	**608**	**405**	**274**	**427**	**406**	**434**	**673**	1019	**637**	712	**514**	**657**
Full Thread	707	764	477	394	496	499	567	1026	1490	991	933	753	875
Siblings	613	730	560	401	590	532	607	804	**1009**	734	759	649	813
Time	584	715	459	460	579	433	542	796	1090	965	776	720	716.05

Here we present scores in terms of micro-averaged F-score (when a gold standard is available for comparison), and by perplexity. A higher F-score indicates a more human like model and a lower perplexity score indicates a less perplexed model. Significance is tested using a Student's two tailed t-test and significant results are quoted when $p<0.01$ (Field, 2013).

4 Results and Discussion

First we will discuss the results from the 3 annotated data sets (1, 2 and 3). Using an F-score metric we find that, for all three annotated sets that grouping comments using the metadata features author and time does not improve topic clustering. Grouping comments using thread based features was more sucessful. We found that combining comments with their replies (the children) and combining comments within the full thread sets significantly beat the standard baseline (Table 2).

The results differ when judged by perplexity (Table 3). We found for two of the comment sets (2 and 3) the children data set gave models that were significantly less perplexed than the standard baseline but this was not the case for comment set 1. For comment set 1 no models beats the baseline using the perplexity metric.

When we look at all of the data, judged using a perplexity score (Table 3), we found that the combined children data sets consistently created models (for 10 out of the 13 sets) that are significantly less perplexed than a standard baseline. For one of the datasets the data combined with other replies to the same message, the siblings set, beats the baseline. For two of the sets no combination method beats the baseline.

The automated results and human results as indicated by perplexity and micro-averaged F-score are not in complete agreement, but there are some commonalities. Both sets of results indicate that the group that combines responses with comments (the children group) has the highest agreement with the human model, and it consistently produces the least perplexed model.

5 Conclusions

It is worth noting that although we focus here on newspaper comments, the need for summarisation applies to any web-based chat forum and the findings therefore have a wide applicability.

LDA topic modelling perfoms better with longer documents. Here we have investigated methods for combining newspaper comments into longer documents in order to improve LDA clustering and therefore provide a strong basis for subsequent comment summarisation. We found that combining comments using features derived from the thread structure of the commenting system was more successful than features from the comments metadata. We found that using a combination of a comment and its children provides 'documents' that produce models that can more accurately classify comments into topics than other document combination methods. It is likely that the method of grouping comments with their direct replies, their children, is the most successful because commentors interact with the other comments through the thread system (rather than newest or oldest first) and they add topically relevent information to the threads. It also indicates that topics in threads evolve, meaning that grouping the entire thread together into a single document works less well than grouping the immediate decendants - the children.

We found that these results were generally consistent, but not identical, across two metrics - perplexity and F-score. We therefore confirm that the perplexity measure is a useful metric in this domain when annotated data is not available.

References

Ahmet Aker, Emina Kurtic, AR Balamurali, Monica Paramita, Emma Barker, Mark Hepple, and Rob Gaizauskas. 2016. A graph-based approach to topic clustering for online comments to news. In *Advances in Information Retrieval*, pages 15–29. Springer.

Alexandra Balahur, Mijail Alexandrov Kabadjov, Josef Steinberger, Ralf Steinberger, and Andres Montoyo. 2009. Summarizing opinions in blog threads. In *PACLIC*, pages 606–613.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. 3:993–1022.

Giuseppe Carenini, Raymond T Ng, and Xiaodong Zhou. 2007. Summarizing email conversations with clue words. In *Proceedings of the 16th international conference on World Wide Web*, pages 91–100. ACM.

Gabor Cselle, Keno Albrecht, and Rogert Wattenhofer. 2007. BuzzTrack: topic detection and tracking in email. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 190–197. ACM.

YaJuan DUAN, CHEN ZhuMin WEIF uRu, ZHOU Ming Heung, and Yeung SHUM. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 763–780.

Andy Field. 2013. *Discovering statistics using IBM SPSS statistics*. Sage.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 40–48. Association for Computational Linguistics.

Derek Greene and James P Cross. 2015. Unveiling the political agenda of the european parliament plenary: A topical analysis. *arXiv preprint arXiv:1505.07302*.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513. Springer.

Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. 2(3).

David Inouye and Jugal K Kalita. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 298–306. IEEE.

Elham Khabiri, James Caverlee, and Chiao-Fang Hsu. 2011. Summarizing user-contributed comments. In *ICWSM*.

Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. 2014. Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science*, pages 161–165. ACM.

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *LREC*, pages 462–468.

Clare Llewellyn. 2016. Guardian Comments data. `http://homepages.inf.ed.ac.uk/s1053147/data/comments_2016.html`. [Online; accessed 09-June-2016].

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 265–274. ACM.

AK McCallum. 2002. MALLET: a machine learning for language toolkit.

Shamima Mithun and Leila Kosseim. 2009. Summarizing blog entries versus news texts. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 1–8. Association for Computational Linguistics.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 105–108. Association for Computational Linguistics.

Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2011. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 685–688. Association for Computational Linguistics.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. 45(4):427–437.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM.

Stephen Wan and Kathy McKeown. 2004. Generating overview summaries of ongoing email thread discussions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 549. Association for Computational Linguistics.