Improving Topic Model Clustering of Newspaper Comments for Summarisation

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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 many different sources. Summarising the content of these comments allows users to interact with the data at a higher level.

The current state of the art within the comment summarisation field is to cluster comments using 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 extracted to construct a summary of the cluster. In this paper we focus on the clustering subtask. It is important that the clustering is optimal as the subsequent extraction relies on a good set of clusters. Research in a similar domain has found that topical mistakes were the largest source of error in summarising blogs (Mithun and Kosseim, 2009).

Comment data, as with many social media datasets, differs from many 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 which are built using longer documents and using more documents are more accurate. It is thought that including documents which are under 50 words is detrimental to model accuracy.

In our corpus the number of comments on each newspaper article is finite and the topics discussed within each set have been inspired by the original article. Therefore, increasing the set with data from external sources is not an option.

In this work, we investigate two subtasks of the clustering task: firstly, we explore whether we can combine comments within a comments dataset to form larger documents to improve the quality of clusters; and secondly we investigate whether we can improve the quality of the clusters by removing comments from the set if they are under a threshold size of 50 words.

2 Methods

2.1 Data

The work reported here is based on comments from news articles taken from the online, UK version of the Guardian newspaper. 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 and adding their comments to any point in the thread.

In this work we refer to all of the comments on a single article as a comment set. Gold standard data was produced by human(s) assigning all comments from a comment set to topic groups. The gold standard data set contained three comment sets. For one comment set two humans assigned groups (3bb88) and for two comment sets (3z4h5, 3z8ht) 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 member. Once these were removed both annotators had created 14 clusters. The human-human F-Score was 0.6066 including the single clusters and 0.805 without. All annotated sets have the clusters with a single members removed. Details can be seen in table 1.

A further 10 sets of comments were collected which were not annotated. Table 2 shows the composition of these comment sets.

<table>
<thead>
<tr>
<th></th>
<th>3bb88</th>
<th>3z4h5</th>
<th>3z8ht</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>160</td>
<td>230</td>
<td>181</td>
</tr>
<tr>
<td>% over 50 words</td>
<td>58</td>
<td>52</td>
<td>29</td>
</tr>
<tr>
<td>Median words</td>
<td>59</td>
<td>54</td>
<td>26</td>
</tr>
<tr>
<td>Mean words</td>
<td>80</td>
<td>81</td>
<td>45</td>
</tr>
<tr>
<td>Min words</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Max words</td>
<td>332</td>
<td>463</td>
<td>541</td>
</tr>
<tr>
<td>Authors</td>
<td>67</td>
<td>140</td>
<td>112</td>
</tr>
<tr>
<td>Threads</td>
<td>54</td>
<td>100</td>
<td>82</td>
</tr>
<tr>
<td>Groups of siblings</td>
<td>126</td>
<td>186</td>
<td>154</td>
</tr>
<tr>
<td>Time Segment</td>
<td>77</td>
<td>113</td>
<td>68</td>
</tr>
<tr>
<td>Human topics</td>
<td>14</td>
<td>21</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 1: Comment Set Composition - Annotated data

<table>
<thead>
<tr>
<th></th>
<th>3yk5j</th>
<th>3yy6n</th>
<th>3mq8g</th>
<th>3jct7</th>
<th>3ycjn</th>
<th>3y58t</th>
<th>3kvze</th>
<th>3jm4q</th>
<th>3yp8a</th>
<th>3m2j5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>51</td>
<td>121</td>
<td>169</td>
<td>176</td>
<td>205</td>
<td>254</td>
<td>328</td>
<td>373</td>
<td>397</td>
<td>661</td>
</tr>
<tr>
<td>% over 50 words</td>
<td>39</td>
<td>37</td>
<td>36</td>
<td>18</td>
<td>38</td>
<td>49</td>
<td>44</td>
<td>26</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>Median words</td>
<td>44</td>
<td>34</td>
<td>34</td>
<td>24</td>
<td>35</td>
<td>49</td>
<td>43</td>
<td>24</td>
<td>25</td>
<td>31</td>
</tr>
<tr>
<td>Mean words</td>
<td>58</td>
<td>53</td>
<td>45</td>
<td>38</td>
<td>69</td>
<td>72</td>
<td>61</td>
<td>40</td>
<td>43</td>
<td>48</td>
</tr>
<tr>
<td>Min words</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Max words</td>
<td>216</td>
<td>365</td>
<td>174</td>
<td>576</td>
<td>775</td>
<td>481</td>
<td>618</td>
<td>543</td>
<td>446</td>
<td>771</td>
</tr>
<tr>
<td>Authors</td>
<td>28</td>
<td>65</td>
<td>105</td>
<td>103</td>
<td>111</td>
<td>120</td>
<td>204</td>
<td>240</td>
<td>246</td>
<td>420</td>
</tr>
<tr>
<td>Threads</td>
<td>21</td>
<td>53</td>
<td>71</td>
<td>67</td>
<td>80</td>
<td>95</td>
<td>132</td>
<td>198</td>
<td>164</td>
<td>319</td>
</tr>
<tr>
<td>Groups of siblings</td>
<td>45</td>
<td>108</td>
<td>139</td>
<td>148</td>
<td>160</td>
<td>205</td>
<td>256</td>
<td>314</td>
<td>320</td>
<td>553</td>
</tr>
<tr>
<td>Time Segment</td>
<td>33</td>
<td>72</td>
<td>110</td>
<td>76</td>
<td>117</td>
<td>142</td>
<td>160</td>
<td>124</td>
<td>119</td>
<td>203</td>
</tr>
<tr>
<td>Automatics topics</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>18</td>
<td>18</td>
<td>16</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Comment Set Composition - Un-Annotated data

2.2 Data Manipulation

Previous work has shown that topic modelling works best when used with longer documents (Tang et al., 2014). Therefore we have investigated two approaches: 1) comparing methods for combining the individual comments into larger ‘documents’ using metadata features; and 2) testing whether the removal of comments that are smaller the 50 words from the training and/or the test set improves performance.

2.2.1 Combining Data

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: 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.

TIME: 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). Comments are grouped together within a ten minute segment.

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 comments is grouped with its siblings, all other comments that reply to a specific comment.

Related comments are combined together, according to the method, to form a single document.

2.2.2 Removing Data
Several studies have been conducted which indicate that topic models built using short documents are less effective. In this experiment we compare topic models trained and tested on three different sets of comments: models trained using comments longer than 50 words and tested with comments longer than 50 words (here after called tested over 50); models trained using comments longer than 50 words and tested with all comments (trained over 50); and one trained and tested with all comments (all).

2.3 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 for which it has the highest score. This approach was implemented using the Mallet tool-kit (Mccallum, 2002).

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 of 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 size. When the data had not been annotated by humans the cluster size 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 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.

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. Therefore the process is repeated 100 times, so that an average score and a standard deviation can be supplied.
2.4 Metrics

There are two main metrics that are provided 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 (Tang et al., 2014).

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).

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).

3 Results and Discussion

3.1 Combining Data

First we will discuss the results from the 3 annotated data sets (3z8ht, 3z4h5 and 3bb88). Using a F-score metric we find that, for all three annotated sets, the combined children and full thread sets significantly beat the baseline standard (fig.1).

When judged using a perplexity score (fig.2), we see that the combined children set is the less perplexed model for two of the comments sets (3z8ht and 3z4h5). No model is less perplexed than the baseline standard in the 3bb88 comment set. Additionally, the baseline is a significantly worse model than the combined author set for the 3z4h5 comment set and, the full thread and time combinations for the 3z8ht comment set.
We can conclude that the automated results and human results as indicated by perplexity and micro-averaged F-score are not in complete agreement, although there are some similarities. 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 the least perplexed model.

Secondly we found for the results for the data which was not annotated (fig.3) that the combined children groups consistently created models (for 9 out of the 10 sets) that are significantly less perplexed than a baseline standard. No other combination method is consistently less perplexed than the standard baseline.

### 3.2 Removing Data

For the removing data task, only the standard and the children combination results are shown in figures 4 and 5. The results are better in the standard set when all comments are used in model building. In the combined children group, building the model with comments which are longer than 50 terms and testing with comments longer than 50 terms provides less perplexed models. Two comments sets achieved best results when training with comments longer than 50 terms but testing with all.

In general the combination of comments with their replies (the children set) increases the length so fewer comments are removed in. This improves this model’s ability to classify the longer comments, but it is not as successful at classifying the comments which are shorter than 50 terms.

### 4 Conclusions

Both metrics and both data sets indicate that using a combination of a comment and children provides ‘documents’ that produce models that can more accurately classify comments into topics than other document combination methods, children make better comment clusters.

We found removing comments that are less than 50 terms increases the ability of a topic model to classify documents that are longer than 50 terms. It does not increase the ability to classify all documents especially shorter documents. This is true with both the standard and children combination set.
References


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

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections to be published in lrec 2014.


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