Reliably Evaluating Summaries of Twitter Timelines

Dominic Rout and **Kalina Bontcheva** and **Mark Hepple**
Natural Language Processing Group, Department of Computer Science
The University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom

Abstract

The primary view of the Twitter social network service for most users is the timeline, a heterogenous, reverse-chronological list of posts from all connected authors. Previous tweet ranking and summarisation work has heavily relied on retweets as a gold standard for evaluation. This paper first argues that this is unsatisfactory, since retweets only account for certain kinds of post relevance. The focus of the paper is on designing a user study, through which to create a gold standard for evaluating automatically generated summaries of personal timelines.

1 Introduction

In 2011, the widely popular microblogging Twitter service had 100 million active users, posting over 230 million tweets a day[1]. The unprecedented volume and velocity of incoming information has resulted in individuals starting to experience Twitter information overload, perceived by users as "seeing a tiny subset of what was going on" (Douglis 2010). In the context of Internet use, previous research on information overload has shown already that high levels of information can lead to ineffectiveness, as "a person cannot process all communication and informational inputs" (Beaudoin 2008).

Automatic methods for ranking and organisation of user timelines are thus required, to enable professional Twitter users to keep up more easily with key work-related information, disseminated via tweets. At the same time, they will allow personal users to focus only on social interactions and content that is truly important to them.

In the context of news and longer web content, automatic text summarisation has been proposed as an effective way of addressing information overload. However, research on microblog summarisation is still in its infancy. Unlike carefully authored news text or scientific articles, microposts pose a number of new challenges for summary generation, due to their large-scale, noisy, irregular, and social nature. This is particularly pronounced in personal Twitter timelines, where many different kinds of posts are mixed together, including strong presence of personal posts (see Section 2).

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]http://www.guardian.co.uk/technology/pda/2011/sep/08/twitter-active-users (Visited May 7, 2012).

The first step of summarising a user timeline is to identify which tweets are of interest to the given user and should thus be included in the summary. In this paper, the prioritisation of content within a user's home timeline is treated as a ranking task. Variants of this problem have been addressed in previous work (see Section 3), although in many cases, the proposed approaches have not been evaluated rigorously, to prove their accuracy on real user timelines. Moreover, research is fragmented, with different types of linguistic and Twitter-specific features evaluated on different data sets. To the best of our knowledge, a thorough, comparative evaluation of the different methods, on a shared data set, has not yet been reported.

The goals of this ongoing work are as follows:

1. To develop a more robust, representative evaluation data set, based on an analysis of the problem of reranking tweets.
2. To explore the hypothesis that there are different possible definitions of saliance within Twitter, all of which may be equally valid.
3. To compare our own work in Twitter content ranking with that of others in a rigorous manner.
4. To validate the hypothesis that the variable nature of tweets is the Achilles' heel of state-of-the-art content ranking for Twitter.

This paper, in particular, examines in detail the first research question.

2 Summarisation of Twitter Timelines

Microblog posts differ significantly from longer textual content, which renders most of the state-of-the-art text ranking and summarisation methods ineffective for ranking Twitter content. For instance, Inouye *et al* (Inouye and Kalita 2011) compared several extractive summarisation algorithms, reporting that centroid and graph-based algorithms do not perform well.

The first challenge comes from the *idiosyncratic, diverse nature* of microposts. Naaman *et al* (Naaman, Boase, and Lai 2010) found over 40% of their sample of tweets were "me now" messages, that is, posts by a user describing what they are currently doing. Next most common were statements and random thoughts, opinions and complaints and

information sharing such as links, each taking over 20% of the total. Less common tweet themes were self-promotion, questions to followers, presence maintenance e.g. "I'm back!", anecdotes about oneself and anecdotes about others. Naaman *et al* group users into *informers* and *meformers*, where meformers mostly share information about themselves.

Therefore, given that a timeline contains so many different kinds of information, an effective summary would require a great deal of diversity, in order to cover most of the key content. Naaman *et al* also report a number of differences between these two kinds of users, which motivates our hypothesis that the ranking and summarisation of user timelines needs to take into account the specific user and what is interesting to them.

The second challenge comes from *limited length* (maximum 140 characters in Twitter). Consequently, a tweet (or a similar micropost) is not easily comparable to a coherent document, which is the classical document-based summarisation task. Even though short units of text have been summarised in the past (e.g. single paragraphs), microposts, such as tweets, are almost unique in their conversational nature, diversity, length, and context-dependence. In addition, while longer web texts may reference one another, references between tweets are often implicit and difficult to identify.

Thirdly, there is great deal of *redundancy* in a Twitter stream, since tweets are contributed by a number of authors. Large news events may garner many independant tweets with very similar themes, and the same article or event could be shared dozens of times. Additionally, retweets of the same content may appear repeatedly within the home timeline, with the user really only interested in reading the post once.

2.1 Task definition

From an algorithmic perspective, the summarisation of personal user timelines consists of four distinct steps:

1. Collect a window of contiguous tweets from a user's home timeline.
2. Separate the contents of that window into discrete textual units.
3. Derive automatically a number of linguistic and other features for each such unit.
4. Use these features to construct a coherent summary of the salient information in that window.

Classical multi-document summarisation makes the assumption that all given documents are about the same topic or event (Nenkova and McKeown 2011). However, as argued above, user timelines are diverse, not just in terms of topics, but are also in terms of interestingness and relevance to the user. Therefore, we argue that the first step towards successful summarisation of topically diverse, personal Twitter timelines is to filter out irrelevant content.

The focus of this paper is on the problem of ranking microposts by interestingness to the given user. Given the home timeline of a Twitter user, containing incoming messages from the people he follows, the goal is to score tweets according to how 'interesting' or 'salient' they are to the particular user. These scores can then be used to filter out the irrelevant tweets.

A sentence is commonly the smallest unit of text that can be considered for inclusion in an extractive summary. Since tweets are very short and rarely contain multiple sentences, a single tweet could be considered equivalent to a sentence. This leads naturally to our formulation of the problem as ranking and filtering entire tweets by relevance.

The task has been described variously as "tweet recommendation" (Yan, Lapata, and Li 2012), "personalised filtering" (Kapanipathi et al. 2011) and "tweet ranking" (Uysal and Croft 2011; Duan et al. 2010). The goal could have been defined in a number of other ways - for example, one might attempt to extract the most useful snippets of conversation from the window of tweets, or the most interesting URLs shared, however we choose to concentrate on the reranking of tweets in the timeline by salience.

In order to make the task more concrete and easier to evaluate with users, we evoke the scenario where a user has been absent from their social network for a full 24 hours. In that case, the perfect summary would contain all of the content that they would still consider relevant and interesting at the end of this period. The use of a whole day is fairly arbitrary, though it importantly imposes a time frame and a more precise experimental setting.

3 Related Work

3.1 Modelling Tweet Relevance

Due to the varied nature of tweet timelines, as well as the complex nature of the Twitter social connections, a micropost in a timeline can be salient for a number of different reasons, as demonstrated in previous work.

Topical relevance is among the most widely studied and is typically specific to the user. Topically relevant tweets are often personal and of interest due to a user's specific location, hobbies or work. Research that captures this kind of relevance has built bag-of-word models for users and used these to score tweets, typically using techniques such as TextRank, LexRank and TF.IDF (Inouye and Kalita 2011; Wu, Zhang, and Ostendorf 2010).

These bag-of-word models can be sparse, so attempts have been more to enrich user interest models with semantic information. (Abel et al. 2011; Kapanipathi et al. 2011; Michelson and Macskassy 2010) create user profiles using semantic annotations grounded in ontologies (e.g. DBpedia) and, in the former case, also expanding URLs mentioned in the user's own posts. None of these approaches, however, have been evaluated formally on timeline summarisation.

Some tweets can also be considered universally relevant, regardless of the interests of the specific user. Such tweets include those that are humorous, or those that originate from trustworthy or influential authors. (Huang, Yang, and Zhu 2011) give prominence to tweets of higher quality, judged both linguistically and in terms of the reputation of the author. Likewise, (Duan et al. 2010) use measures of user

authority in order to rank search results; their work also demonstrates the use of RankSVM in order to learn rankings from features attached to tweets.

The social relationship between an author and the reader can also affect relevance. (Yan, Lapata, and Li 2012) present personalised recommendations for tweets, incorporating both textual and social features using a graphical model. Similarly, (Chen et al. 2010) recommend URLs to Twitter users based on both previous posts and on social connections.

This paper extends current understanding of tweet relevance, by presenting an experimental analysis of the reasons users find Twitter content in their timelines interesting (see Section 5.2). We hypothesise that there are many types of salience for tweets. Consequently, a tweet filtering or ranking approach should either specifically address one kind of relevance but acknowledge that others exist, or attempt to account for all of them, if possible. A system that is to perform effectively on naturally occurring tweet timelines, evaluated against a user-annotated gold standard, must adopt the latter approach.

3.2 Summary Evaluation

Quantitative evaluation of text-based summaries is a challenging task, since summaries can take many different forms and different people would summarise the same (set of) documents differently (Nenkova and McKeown 2011). The problem can be even more acute in Twitter, where summaries may take the form of lists of tweets, or visual representations of tweet volume or other types of timeline.

Additionally, since it is generally impossible to read all of the tweets for a corpus in order to summarise them, human generated summaries for comparison can be very incomplete or may miss information which should have been included.

For textual summaries of relatively small and topically coherent collections of tweets, researchers have used the standard DUC evaluation measures. For instance, (Sharifi, Hutton, and Kalita 2010) gather 100 tweets per topic for 50 trending topics on Twitter, and then ask volunteers to generate summaries which they feel best describe each set of tweets. The automatic summaries produced by their algorithm are then tested using a manual content annotation method, evaluating whether or not the same content is available as in the gold standard, and automatically using Rouge-1. Similarly, (Harabagiu and Hickl 2011) make use of model summaries, although they eschew Rouge in favour of the Pyramid method (Nenkova and Passonneau 2004).

Another class of evaluation treats summarisation as a document retrieval problem. (Chakrabarti and Punera 2011) use knowledge about the significant plays in a sports match to build a gold standard for what must be discussed in a match summary. After simplifying the problem somewhat by assuming that the search process is perfect and manually validating the input to their system, they calculate recall based on the events from the game, before asking users to subjectively classify the content of tweets in the summaries. Tweets with a certain content type are said to be irrelevant, providing a measure of precision.

(Becker et al. 2011) similarly ask users to score the quality, relevance and usefulness of returned tweets, although they do not attempt to address the problem of calculating recall.

Tweet ranking methods have used retweets as a gold standard for relevance (Yan, Lapata, and Li 2012; Uysal and Croft 2011). However, this methodology is somewhat problematic, in that it requires an explicit, observable feature which is closer to a user recommending the content (e.g. Facebook's like button) than them actually retweeting it. A user could conceivably like a tweet but not be aware of anyone amongst their followers who would enjoy it, in which case they would not retweet. Additionally, they may retweet content to their followers that they are personally not interested in, as a way of building social capital.

An alternative to gathering retweets as judgements against which to evaluate a system is to ask users to indicate which of the tweets in their timelines are interesting to them. We will discuss this approach to evaluation in the rest of this paper. Such judgements were gathered from volunteers by (Ramage, Dumais, and Liebling 2010) for Twitter, though users were asked to judge tweets from one user at a time, and by (Paek et al. 2010) for Facebook, where users were instructed to rate general content from their feeds.

4 Constructing a Gold Standard

As argued above, retweets are an imperfect proxy for salience. A user is far more likely to be interested in a tweet that they have retweeted, though the reverse is not necessarily true. Users can be interested in content that they do not retweet. There are certain types of tweet that may be interesting but will almost never get retweeted, for example tweets that are part of conversations. It would not make sense to broadcast the important parts of a personal discussion to all of one's followers.

As an alternative to using retweets as an indicator of relevance, we approach active twitterers directly, and carry out a user study in which they are asked to make explicit judgements of which tweets they find relevant. Specifically, Twitter users are asked to indicate the relevance of 100 consecutive tweets from their own timeline (using a binary judgement), and these judgements are stored as a gold standard. A system that scores tweets according to their 'interestingness' for the given user can then be evaluated by comparing its rankings with those contained in the gold standard.

Notably, out of the 585 positive judgements in an early version of our gold standard, only 2 were retweeted by the participants themselves. The Twitter API provides incomplete information about who retweeted what, however, so the figure may in fact by higher than this. Nonetheless, only 66% of 'interesting' tweets were retweeted by anyone at all, in the whole of Twitter. This demonstrates empirically that validation based on retweets is incomplete.

In order to better capture salience, we propose to use a collection of manual judgements from volunteer users as a gold standard for evaluating rankings of twitter home timelines. This approach however, is not without some limitations, namely that users cannot be expected to review large portions of their timeline, so only a subset of interesting

tweets can be annotated. A gold standard constructed in this way will nontheless allow systems to be evaluated more accurately, and will also capture more types of salience as discussed in Section 3.1, allowing these to be addressed in future work.

4.1 User Study

We collected our data set by recruiting 71 Twitter users and asking them to complete a very informal annotation task. The study was advertised via Twitter and Facebook and on a University mailing list. As a result, most participants in the study had a relatively high degree of literacy, and recieved tweets largely in formal English. Several of the users also recieved and sent tweets in languages other than English. These tweets have been left in the dataset, although the scope of this study does not include discovering recommended tweets in e.g. Finnish for users that only tweet in English; this would be a difficult cross-language task.

After agreeing to take part in the study, the volunteer twitterers authenticated a simple web application to access their account. Their most recent 800 incoming tweets were downloaded and stored. Additionally, the first 100 of these were presented to the user for annotation (see Figure 1). They were then asked to read the feed and select the most 'interesting' tweets. By default, all tweets were marked as uninteresting and users were asked to select those that stand out.

In the early stages of our investigation we did not have any information about the kinds of tweets that users would find interesting. The annotation interface was modelled to resemble the original Twitter interface as closely as possible, in order to minimise the possibility that our experiment design could affect their behaviour. The interface also allowed us to gain some insight into the reasons why users would be interested in certain content. To reenforce this, we also allowed users to add an optional free comment to each judgement, explaining the reasons why they selected a tweet as interesting. About 5.8% of tweets were commented on by the participants.

Tweets were displayed in reverse chronological order and users could skim over them, lingering on and marking as interesting only the ones that 'grabbed' their attention. The usernames of their friends were kept instact, and also the thumbnail image from their friends profile was shown. Links were expanded and made clickable, if desired. The target link content appeared in a second window to allow users to return to the study easily.

We did not place any limit on the number of tweets that could be marked as interesting, or specify a lower bound. As a result of this lack of more strict guidelines, different users marked vastly differing portions of their streams. Some users selected only 3 tweets, while others selected as much as 50%. We believe that this variation in the number of interesting tweets is caused more by the imprecise guidelines given to the users, rather than by an actual difference in the kind of useful content. Either way, it is difficult to estimate reliably the usefulness of any algorithm when the size of the gold standard varies enormously, since predicting the relevant 50% of tweets is naturally a much 'easier' task than predicting 5%. The effect of this size variation on the evaluation of a simple TF.IDF method is shown in Figure 2.

4.2 Scoring approaches

The judgements created by users are binary (i.e. relevant and not relevant). However, the summariser must generate scores for relevance for the input tweets, in order that high scoring examples can be used to construct the summary itself. This is a common situation in information retrieval, where for example search results need to be ranked but the gold standard is constructed from binary information, i.e., whether or not a user actually clicked a result in response to a given query.

For our initial evaluation, we choose to use the mean average precision (MAP) score to represent the quality of a tweet scoring algorithm (Yue et al. 2007). This quantitative measure was also adopted by (Ramage, Dumais, and Liebling 2010; Yan, Lapata, and Li 2012).

Other work has compared algorithms using normalised discounted cumulative gain (nDCG) (Yan, Lapata, and Li 2012; Huang, Yang, and Zhu 2011; Duan et al. 2010) and we plan to also calculate this metric in future work (Järvelin and Kekäläinen 2000).

Even though both methods can be used on sets of relevant documents of varying sizes, when that set varies from a small percentage of the candidates up to 50% or more as in our data, calculated scores will naturally vary hugely from one query to the next. On our tweet gold standard, this leads to a dramatically different performance score for each user, such that statistically significant performance improvements are very hard to reach. The relationship between the size of the relevant set and the performance attained by a simple method is shown in Figure 2.

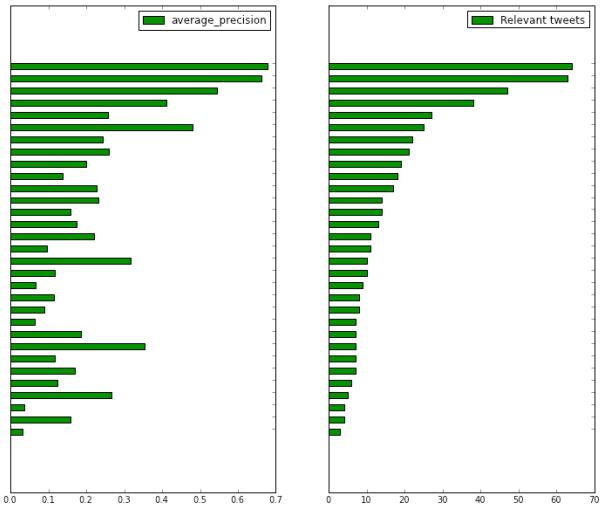

Figure 2: Average precision for collections with varying counts of relevant tweets

4.3 Temporal effects

Despite the apparent problems with the gold standard, we implemented a number of baselines against which other

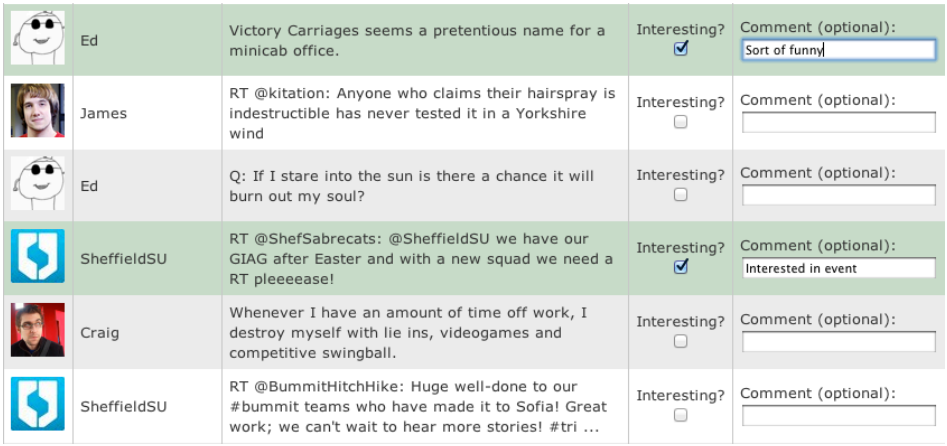

Figure 1: Interface used to annotate interesting posts. Selected tweets are highlighted

methods could be evaluated. One such baseline was random tweet selection. Another was the temporal baseline, which ranks tweets reverse chronologically. On our dataset the temporal baseline gave dramatically higher performance than the random one. In contrast, the TF.IDF method, which has been shown to outperform baselines (Ramage, Dumais, and Liebling 2010; Inouye and Kalita 2011), gave results which were lower than the temporal baseline (Cosine TF.IDF: 0.19 MAP, Temporal: 0.22 MAP).

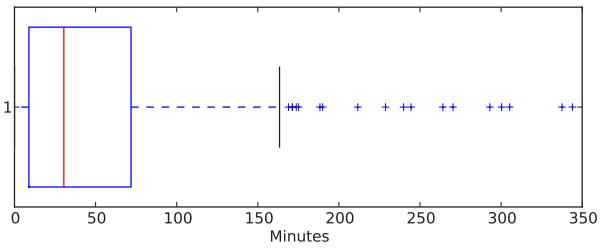

Figure 3: Distribution of 240 positive judgements in difference between posting and time of annotation. Outliers above 6hrs (8% of points) have been removed for readability. Actual upper limit is 48hrs.

This motivated us to analyse the position of the user-marked interesting tweets in the reverse chronological sample of 100 tweets, which was annotated by each user. Given Twitter's streaming nature, temporal effects suggest that perhaps more recent tweets should naturally be of greater interest to the user. However, the effect of this was much stronger than anticipated, with users much more likely to choose tweets towards the very start of the list, as shown in Figure 3. This is especially surprising since the time span covered by the 100 consecutive tweets is not that large. Figure 4 shows in more detail the distribution of the length of time covered by samples in our study. Most candidate sets for annotation only covered a time period of less than an hour.

There we hypothesize that the temporal effect is not as strong in reality as the data from this study would suggest. Certainly, in the work of (Ramage, Dumais, and Liebling 2010) a temporal baseline did not outperform random ordering by as big a margin as on our data. Instead, we argue that the skew is a result of the lack of more strict annotator guidelines in our study, which allowed users to select as interesting vastly different numbers of tweets. This variance creates a situation in which incredibly dramatic leaps in performance would be required to achieve statistically significant improvement in results.

Another possible explanation for this strong temporal bias in interestingness is annotator boredom, coupled with repetitive tweets in the user timeline (the redundancy discussed above). This leads us to believe that asking users to annotate 100 tweets in one session may have been too great a number.

5 Redesigned user study

In order to address the design weaknesses in the first version of the study, we plan to carry out a redesigned corpus collection experiment within the next few months.

In order to eliminate the variance in the number of tweets selected as interesting by the different users, volunteers in the new study will be asked to select a fixed number of 8 tweets as interesting, out of a sample of 50. This is similar to the word-based restriction on summary sizes, commonly used in evaluation initiatives for multi-document news sum-

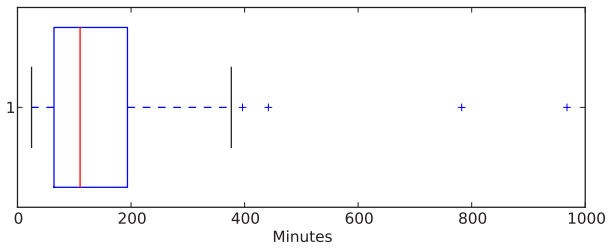

Figure 4: Distribution of time in minutes between first and last tweet in a collected sample.

marisation, e.g. the TAC Guided summarization task[2].

Whilst restricting the number of relevant tweets to 1 in 6 does force conformity onto participants and limits our dataset, we believe that this is a reasonable simplification to make and one that will make evaluation of prospective approaches far more straightforward. This ratio was arrived at empirically, as the mean percentage of tweets in our initial study that were annotated as relevant by users.

The relatively small window of tweets to be judged will help to partly mitigate the problem of redundancy - where users are unlikely to select two tweets that have similar content - because such pairs are less likely to occur. We will also address the issue by automatically removing retweets from the candidate sets.

In addition, users were far more likely to pick a tweet higher up in timeline, which coupled with its reverse chronological ordering, introduced a strong temporal bias in the collected data. The next study will instead order tweets in a random fashion, instead of trying to mimic existing Twitter reading tools. We hope that by observing the temporal effect once tweets have been ordered randomly, we can effectively model the relationship between the age of a tweet and its likelihood of still being relevant.

One problem with reordering tweets is that it may break the flow of conversations and dialogs, thus creating posts which are difficult to read due to the removal of their surrounding context. We plan to address this by changing the web interface, to allow users to expand the context surrounding a tweet by clicking on it. This will display either a few earlier tweets from the same author, or a chain of replies if the tweet forms part of a dialog.

The new study will also employ two different strategies for selecting the sample of 50 tweets to be annotated by the volunteers. Both strategies will allow tweets to be sampled from a 24 hour period, collected from the participant's own home timeline. The first sampling method will choose tweets from the candidate set at random. The second sampling method (detailed in Section 5.1) will draw the sample from topically coherent tweets instead. The sampling method used for a given user will be alternated to allow for a balanced comparison. In both cases, volunteers will be asked to choose the 8 most interesting tweets from a sample of 50, resulting in a 16% summary compression rate. Using a ratio much larger than this might render the task too simple, while a smaller ratio would make the number of positive examples too small, to enable comparative evaluation of different methods and features. The sample size will be halved down to 50, since we observed in the first study that when as many as 100 tweets were displayed, users were more likely to annotate only the earlier ones.

5.1 Creating coherent topics

A distinguishing characteristic of this research is in its focus on heterogenous user timelines. We plan to develop salience measures capable of capturing tweet variability in terms of type and purpose (the latter being the focus for future work).

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

In order to allow a more in-depth understanding of what makes tweets interesting to a given user, the second study will ask users to group their tweets according to some criteria, prior to drawing a sample of 50 to be annotated. In this way, we aim to validate the hypothesis that the major weakness of current tweet summarisation systems is caused by the inter-twined mixture of topics and conversations in tweet timelines, which are currently ordered purely on a temporal basis.

Participants will be shown first a tag cloud representation of their Twitter timeline. This view is naturally imperfect - containing noisy irrelevant terms and not necessarily promoting the most indicative ones. However, by asking users to select terms from the cloud (thus indirectly selecting the matching tweets), we can create a sample of tweets that is to some extent topically coherent. Even if the coherence is not very strong, we hope to demonstrate that the enormous entropy in user timelines is at least partly holding back the performance of state-of-the-art methods.

The tag cloud interface will allow users to select several terms and the tweet sample will be drawn from the union of tweets containing their selections. In cases when there are fewer than 50 tweets containing the chosen terms, participants will be asked to keep selecting more terms until the sample is complete. If there are no more related terms left, the remaining tweets will be drawn at random from the rest of the timeline.

Tag clouds will be built from single terms and usernames taken from the home timeline. We determined through informal experimentation that a combination of an IDF index and weighting according to the frequency with which the participants themselves use those terms or interacted with those users produces useful, if somewhat noisy, tag clouds. Other schemes including named entity extraction were considered.

5.2 Annotating Reasons for Salience

Since one of our goals is to identify and classify the different reasons that make tweets interesting to a given user, the second user study will also gather more detailed rationale information from the volunteers. As mentioned above, the first version of the study asked users to supply optional 'Other Comments' alongside their interestingness judgements. This was well received and was used by the volunteers to explain their decision, with roughly 5.8% of the judgements in our study being accompanied by comments.

Stated reasons included agreement with the content, liking the products and services mentioned, being a fan of the author (e.g. a celebrity) or knowing them personally. A manual analysis of these textual responses was carried out, which enabled us to generalise them into a classification, describing different types of tweet relevance.

Having generalised the responses from the first version of our study, we formed 7 salience reasons, which were validated by asking 6 of volunteers to apply them to a set of 100 responses. A Fleiss' Kappa score of 0.53 was calculated for the set indicating moderate agreement. The labels used and the proportion of comments assigned to that label (averaged across annotators) are shown in table 5.2.

Positive Comments	
Relevant to work/location/interests	29%
Funny/Clever observation	20%
Contains interesting or surprising information	18.7%
Sympathise/agree with the sentiment	6%
Mentions something they like	6%
They like or trust the author	4%
Addresses reader or a friend directly	4%
Other	12.3%

We offered an 'other' category during annotation which was used in cases where it was unclear what the author meant by their comment. The portion of responses labelled 'other' is also shown in the table. Overall we feel that the categories here are exhaustive enough to be used in our ongoing work.

These classes demonstrate the varied reasons for which a tweet might be interesting. Some of them are very personal, requiring in-depth knowledge of the user's location, work, interests and relationshiops to predict. Others are less so - humour and surprise are arguably less personal.

This newly developed classification will be offered to participants in the second study, who will be asked to explain their decisions by selecting interestingness reasons from this list. An open-ended, text feedback field will still be shown at the bottom, to allow users to enter new types not already covered by the current classification. Based on this more detailed data, we aim to analyse and quantify more precisely the reasons that make a tweet relevant or irrelevant to a given user.

5.3 Participation

This work relies heavily on user participation in order to generate a data set, large enough to be useful for evaluation. An even larger sample would be needed, in order to create a training set for supervised machine learning. The first version of our study had 71 respondents in total. Each completed the specified annotation task just once.

Since it is not possible to measure inter-annotator agreement (IAA) for this summarisation task, we will ask participants to complete the annotation task several times, on samples drawn in different ways (random versus topic-focused), over a period of several weeks. In this way, more gold-standard data will be gathered for development and evaluation. In addition, since an individual's preferences are unlikely to change between iterations, clearer measures of significance should be attainable.

The new study will be advertised, as before, using a combination of Twitter and University mailing lists. Repeat participation will be solicited via email and automated tweets. We hope this time to reach a wider audience, as the initial version of the study was completed late in the year, coinciding with a number of staff and student deadlines.

5.4 Limitations of Study Design

It is common practice in text summarisation to construct multiple model summaries for the same document set using several annotators. However, this approach is not suitable for ranking tweets in personal timelines, since it is likely that every user will read their timeline differently, looking for slightly different information. Indeed, it is useful for a system to attempt to operate in a general enough fashion by making fewer assumptions about user behaviour. As such, our task is to generally advance the state of the art in Twitter summarisation by designing methods that perform well for most users in the gold standard.

Our study still relies on real timelines of real Twitter users. Even though we assume that the new more restricted and longitudinal design will lead to more representative and realistic judgements, the volume of tweets to be annotated is likely to be relatively small, limited by the number of participants. However, widening our participant sample is not trivial. This type of in-depth, personal annotation is not well suited to popular crowdsourcing platforms such as Mechanical Turk or Crowdflower, since no measure of agreement can be used to eliminate spam responses. This weakness could perhaps be addressed by constructing model timelines for artificial Twitter users and having those judged by many unknown annotators. It would be difficult, however, to prove that the model timelines are representative of real user behaviour.

Participants in the user study could be given a selection of tweets and asked to pick those that they think best summarise the entire set. Such task definition might make it feasible to assign the same timeline to several users, however it assumes that the set can be read objectively, and while such an assumption may be valid for a collection representing an event, for example, it is unlikely to hold for a user's personal timeline. For the latter, we believe that salience is simply far too subjective and contextual to have multiple annotators for the same collection.

6 Conclusion

By addressing some of the weaknesses in the evaluation of state-of-the-art Twitter summarisation, we hope to not only provide fairer, more representative results for the perfomance of such systems, but to allow a wider range of approaches to be developed. This, however, necessitates the creation of a gold standard for tweet summarisation, since, as demonstrated in this paper, a gold standard based on retweets is unsatisfactory. While retweets are relatively easy to collect in large numbers, they only account for certain kinds of post relevance.

This paper discussed the first attempt at collecting a gold standard of interesting tweets from user timelines. Shortcomings in experimental design resulted in an imbalanced dataset, which motivated us to design a follow up data collection experiment, which is currently under way. The resulting data set will be made available by request from the first author.

In future work, the newly collected gold standard will be used for development and evaluation of automated methods for summarisation of personal timelines. We plan to explore novel features and address the heterogeneity of the timelines in our future work.

One potential avenue for future work is in relevance feedback, allowing users to express which of the tweets in an au-

tomatically generated summary they are most interested in. These selected tweets may form a small but exact corpus of examples of content in which a user is very interested. In this case, the evaluation measures used would need to be reconsidered, representing the increase in performance as more data becomes available.

Some tweets are interesting to readers because of a social relationship that they have with the poster. Though we can easily gather information about the social networks of participants in our study, we do not have any gold standard information about the types and history of their relationships. This may be gathered at a later date.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1). We would also like to thank the reviewers for their detailed comments and suggestions, which helped us improve this paper significantly.

References

Abel, F.; Gao, Q.; Houben, G.-J.; and Tao, K. 2011. Semantic enrichment of twitter posts for user profile construction on the social web. In Antoniou, G.; Grobelnik, M.; Simperl, E. P. B.; Parsia, B.; Plexousakis, D.; Leenheer, P. D.; and Pan, J. Z., eds., *ESWC (2)*, volume 6644 of *Lecture Notes in Computer Science*, 375–389. Springer.

Adamic, L. A.; Baeza-Yates, R. A.; and Counts, S., eds. 2011. *Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM 2011 Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.

Beaudoin, C. 2008. Explaining the relationship between internet use and interpersonal trust: Taking into account motivation and information overload. *Journal of Computer Mediated Communication* 13:550—568.

Becker, H.; Chen, F.; Iter, D.; Naaman, M.; and Gravano, L. 2011. Automatic identification and presentation of twitter content for planned events. In Adamic et al. (2011).

Chakrabarti, D., and Punera, K. 2011. Event summarization using tweets. In Adamic et al. (2011).

Chen, J.; Nairn, R.; Nelson, L.; Bernstein, M. S.; and Chi, E. H. 2010. Short and tweet: experiments on recommending content from information streams. In Mynatt, E. D.; Schoner, D.; Fitzpatrick, G.; Hudson, S. E.; Edwards, W. K.; and Rodden, T., eds., *CHI*, 1185–1194. ACM.

2010. *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics Proceedings, HLT-NAACL, June 2-4, 2010, Los Angeles, California, USA*, The Association for Computational Linguistics.

Douglis, F. 2010. Thanks for the fish but i'm drowning! *IEEE Internet Computing* 14(6):4–6.

Duan, Y.; Jiang, L.; Qin, T.; Zhou, M.; and Shum, H.-Y. 2010. An empirical study on learning to rank of tweets. In Huang, C.-R., and Jurafsky, D., eds., *COLING*, 295–303. Tsinghua University Press.

Harabagiu, S. M., and Hickl, A. 2011. Relevance modeling for microblog summarization. In Adamic et al. (2011).

Huang, M.; Yang, Y.; and Zhu, X. 2011. Quality-biased Ranking of Short Texts in Microblogging Services. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 373–382.

Inouye, D., and Kalita, J. K. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *SocialCom/PASSAT*, 298–306. IEEE.

Järvelin, K., and Kekäläinen, J. 2000. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, 41–48.

Kapanipathi, P.; Orlandi, F.; Sheth, A. P.; Passant, A.; and Passant, A. 2011. Personalized filtering of the twitter stream. In de Gemmis, M.; Luca, E. W. D.; Noia, T. D.; Gangemi, A.; Hausenblas, M.; Lops, P.; Lukasiewicz, T.; Plumbaum, T.; and Semeraro, G., eds., *SPIM*, volume 781 of *CEUR Workshop Proceedings*, 6–13. CEUR-WS.org.

Michelson, M., and Macskassy, S. A. 2010. Discovering users' topics of interest on twitter: a first look. In Basili, R.; Lopresti, D. P.; Ringlstetter, C.; Roy, S.; Schulz, K. U.; and Subramaniam, L. V., eds., *AND*, 73–80. ACM.

Naaman, M.; Boase, J.; and Lai, C.-h. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, 189–192. ACM.

Nenkova, A., and McKeown, K. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval* 5(2-3):103–233.

Nenkova, A., and Passonneau, R. J. 2004. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, 145–152.

Paek, T.; Gamon, M.; Counts, S.; Chickering, D. M.; and Dhesi, A. 2010. Predicting the importance of newsfeed posts and social network friends. In Fox, M., and Poole, D., eds., *AAAI*. AAAI Press.

Ramage, D.; Dumais, S. T.; and Liebling, D. J. 2010. Characterizing microblogs with topic models. In Cohen, W. W., and Gosling, S., eds., *ICWSM*. The AAAI Press.

Sharifi, B.; Hutton, M.-A.; and Kalita, J. K. 2010. Summarizing microblogs automatically. In *HLT-NAACL* (2010), 685–688.

Uysal, I., and Croft, W. B. 2011. User oriented tweet ranking: a filtering approach to microblogs. In Macdonald, C.; Ounis, I.; and Ruthven, I., eds., *CIKM*, 2261–2264. ACM.

Wu, W.; Zhang, B.; and Ostendorf, M. 2010. Automatic generation of personalized annotation tags for twitter users. In *HLT-NAACL* (2010), 689–692.

Yan, R.; Lapata, M.; and Li, X. 2012. Tweet recommendation with graph co-ranking. In *ACL (1)*, 516–525. The Association for Computer Linguistics.

Yue, Y.; Finley, T.; Radlinski, F.; and Joachims, T. 2007. A support vector method for optimizing average precision. In Kraaij, W.; de Vries, A. P.; Clarke, C. L. A.; Fuhr, N.; and Kando, N., eds., *SIGIR*, 271–278. ACM.