Chapter 14

Summarisation of UGC

Dominic Rout

University of Sheffield

Kalina Bontcheva

University of Sheffield

14.1 Introduction

The phenomenal growth of the social Web 2.0 and User-Generated Content (UGC), such as blogs and online social networks, is driven by tapping into the social nature of human interactions, making it possible for people to gain a wider audience for their opinions, become part of a virtual community and collaborate remotely. Engaging actively with this high-value, high-volume, distributed and dynamic content has now become a daily challenge for both organisations and ordinary people. Automating this process through intelligent information access methods is therefore necessary; it is also computationally viable.

Information access to UGC is an emerging research area, combining methods from many fields, e.g. speech and language processing, social science, machine learning, personalisation, and web science. Traditional search methods

are no longer able to address the more complex information seeking behaviour in UGC, which has evolved towards sensemaking, learning and investigation, and social search [74].

At the same time, research on text summarisation of UGC is still in its infancy, especially with respect to aiding information interpretation. Unlike carefully authored news text or scientific articles, UGC poses a number of new challenges for summary generation (see Section 14.3), due to its large-scale, noisy, irregular, and social nature.

We start by providing a brief overview of automatic text summarisation, as originally developed for well-formed, long textual documents, such as news articles and scientific papers. The chapter then surveys the state-of-the-art in UGC summarisation, including research on text-based UGC summaries (Section 14.4), structured, sentiment-based summarisation (Section 14.5), keyword- and topic-based summarisation (Section 14.6), and summary-based visualisations (Section 14.7). Section 14.8 discussed evaluation challenges and methods, followed by a concluding discussion on outstanding challenges (Section 14.9).

14.2 Automatic Text Summarisation: A Brief Overview

Textual summaries are extremely common in online and printed media. Table 14.1 lists some examples of types of summaries which are commonly encountered. These classes have been described as either critical, informative or indicative, where critical summaries attempt to appraise and evaluate a work in some way, informative summaries try to capture the content of the original document, and indicative summaries aim to enable a reader to decide whether or not to read the whole document [56, 65].

Summary	Purpose
Journal abstract	Indicative
News report	Informative
Movie review	Critical
Novel blurb	Indicative
Football highlights	Informative

TABLE 14.1: Example classes of summary

Researchers have developed methods for summarising single and multiple documents, either by combining sentences and phrases extracted from the original texts (called *extractive summarisation*, e.g. [80, 7]) or by using Natural Language Generation (NLG) to create *abstractive*, interpretative summaries (also called concept-to-text generation) (e.g., [11, 94]). The former type of

summary, the extract, is composed entirely from text which can be found in the original document(s). Overall, extractive summarisation is arguably easier to explain and implement, however, the resulting summaries reflect strongly the original document(s), which could be problematic on short texts, such as tweets.

Extractive summaries are generally produced according to two steps.

1. Score textual units (sentences, phrases, paragraphs etc) according to some representation of the document or document set.
2. Generate summaries by selecting high scoring textual units until some desired compression ratio has been achieved.

The textual unit to be included in a summary could be a word, phrase, sentence or whole paragraph depending on the application. Different methods for scoring textual units have been developed, including word frequencies (TF.IDF), sentence position in the document [54], and centroid-based methods [77].

On the other hand, abstractive summarisation algorithms tend to be much more complex. The advantage of the abstractive approaches is that they enable succinct summaries of the content, independent of its original presentation in the source document collection [15], as well as the generation of different (personalised) summaries from the same formal input [11].

In addition to being described by their form (abstractive or extractive) and their purpose (critical, informative or indicative) summaries can also be classified by whether they are derived from single or from multiple documents. The two types are considered somewhat separate, because multi-document summarisation must address different challenges to single document summarisation such as repeated text between documents, order of publishing and inter-document references.

Summaries may be topic-centric (generic), user-focused (i.e. personalised) or query focussed. The former class of summary is meant simply to summarise the content with no bias as to whom can benefit from it. Query focussed summarisation, on the other hand, involves building a summary to meet a specified information need; in this sense it is related to the task of question answering. User focussed or personalised summarisation must model in some way the information needs and interests of a specific user, arranging a summary containing details which they alone may find salient.

14.3 Why is User Generated Content a Challenge?

State-of-the-art automatic text summarisation algorithms have been developed primarily on news articles and other carefully written, long documents

[56, 65]. In contrast, user generated content tends to be very different: often short, strongly grounded in context, temporal, noisy, and full of slang.

In more detail, a study comparing Twitter and New York Times (NYT) news [97] has identified three kinds of topics: event-oriented, entity-oriented, and long-standing topics. Topics are also classified into categories, based on their subject area. Nine of the categories are those used by NYT (e.g. arts, world, business) plus two Twitter-specific ones (Family&Life and Twitter). Family&Life is the most predominant category on Twitter (called 'me now' by [62]), both in terms of number of tweets and number of users. Automatic topic-based comparison showed that tweets abound with entity-oriented topics, which are much less dominant in traditional news media.

With respect to message content, Naaman *et al* [62] 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. Messages posted from mobile devices are more likely to be "me now" messages (51%). Females post more "me now" messages than males. A relatively small number of people undertake information sharing as a major activity; users can be grouped into *informers* and *meformers*, where meformers mostly share information about themselves. Informers and meformers differ in various ways. Informers tend to be more conversational and have more contacts.

These idiosyncratic characteristics of user-generated content are also opportunities for the development of new text summarisation approaches, which are better suited to media streams:

Short messages (microtexts) : Twitter and most Facebook messages are very short (140 characters for tweets). Text summarisation methods could supplement these with extra information and context coming from embedded URLs and hashtags[1].

Noisy content : social media content often has unusual spelling (e.g. 2moro), irregular capitalisation (e.g. all capital or all lowercase letters), emoticons (e.g. :-P), and idiosyncratic abbreviations (e.g. ROFL, ZOMG). Spelling and capitalisation normalisation methods have been developed [37], coupled with studies of location-based linguistic variations in shortening styles in microtexts [35].

Temporal : in addition to linguistic analysis, user generated content lends itself to summarisation along temporal lines, which is a relatively under-researched problem. Addressing the temporal dimension of UGC is a

[1] A recent study of 1.1 million tweets has found that 26% of English tweets contain a URL, 16.6% – a hashtag, and 54.8% contain a user name mention [16].

pre-requisite for summaries of conflicting and consensual information as well changes in opinions over time.

Social context is crucial for the correct interpretation of UGC. Summarisation methods could make use of social context (e.g. who is the user connected to, how frequently they interact), in order to tailor the content of summaries accordingly, e.g. make more prominent content from users of specific interest, e.g. highly authoritative users, groups of similar users.

User-generated : unlike traditional web content, UGC is a rich source of explicit and implicit information about the user, e.g. demographics (gender, location, age, etc.), interests, opinions. This enables new kinds of summaries, e.g. location-based, by age groups.

Multilingual : UGC is strongly multilingual. For instance, less than 50% of tweets are in English, with Japanese, Spanish, Portuguese, and German also featuring prominently [16]. Unfortunately, text summarisation methods have so far mostly focused on English, while low-overhead adaptation to new languages still remains an open issue. Automatic language identification [16, 5] is an important first step, allowing applications to first separate social media in language clusters, which can then be processed using different algorithms.

In our experience, amongst all kinds of user generated content, tweets and similar short status update messages pose the biggest challenge. This is due to the fact, that state-of-the-art text summarisation methods tend to make certain assumptions, which clearly do not hold for the short, interconnected, and noisy messages typically found on Twitter, Facebook, and other similar sites:

- Frequency-based methods, such as TF.IDF must necessarily make the assumption that salient terms will be repeated within a document. However, individual UGC messages, e.g. tweets, are unlikely to contain any repeated terms, so Term Frequency must be defined in some other scope - usually by broadening the definition of a 'Document' to include many posts. Additionally, the IDF for infrequent and out-of-vocabulary words is usually very high, and such terms are very common within microposts.
- Position is difficult to extract from a set of microposts - since messages are generally only ordered temporally, it is not necessarily clear whether or not the ordering within a stream of messages is due to some conscious effort by the author. The assumption that a series of 'documents' extracted from microposts will contain some common structure is somewhat difficult to defend.
- Deeper parsing and discourse analysis methods generally assume that the necessary tools have been developed for a domain. However, parsing

tweets has been shown to be somewhat more difficult than parsing many other kinds of text [79].

Consequently, the appropriateness of using a method developed for traditional text summarisation, without adaptation specific to microposts in particular, is deeply questionable, for a number of reasons.

Firstly, we dispute that a single tweet (or a similar micropost) is in any way comparable to a single document as used in the classical document-based summarisation setting. Though short units of text have been summarised in the past, such as single paragraphs in isolation, short messages, such as tweets, are almost unique in their conversational nature, diversity and length. Additionally, tweets are often far more context-dependent than longer documents. While larger texts may reference one another, references between tweets are often implicit and difficult to identify.

An alternative to defining a document as a single tweet is to form some collection of them and treat that as a single document for summarisation. There are also problems with this approach; the collection will carry far less coherence than a longer document composed by a single author. If automatically collected, the tweets are unlikely to all cover the same specific topic, creating an additional summarisation challenge.

Aside from defining the granularity of a document when summarising tweets, it is also unclear how large a textual unit should be. In some sense, a single tweet might be analogous to a sentence, since their shortness makes messages containing multiple sentences uncommon, but this cannot always be assumed. However, while individual tweets are less self-contained than documents, they are more so than sentences. They do not always 'make sense' when viewed alone, but a collection of tweets embodies less narrative than a paragraph of sentences.

Depending on the source of messages used, Twitter can be informationally efficient or contain lots of redundancy. Tweets can involve lots of co-reference or none at all. Sometimes order matters and sometimes it does not. The problem of defining what summarisation for Twitter actually requires, or what form summary takes, is still very much under-explored. However, the sheer quantity of information nevertheless motivates work in this area.

The rest of this chapter provides an overview of the state-of-the-art in summarisation of different kinds of user-generated content and then concludes with open issues and future work.

14.4 Text Summarisation of UGC

Research on text-based summarisation of user generated content has predominantly focused on three kinds of UGC: online reviews, blogs, and short

status updates (mostly Twitter messages). Both online reviews and blog posts tend to be longer, which makes "traditional" text summarisation techniques easier to apply, including single document and multi-document summarisation methods. In contrast, the summarisation of short status updates has proven much more challenging for state-of-the-art text-based methods. Now let us examine these in more detail.

14.4.1 Summarising Online Reviews

Summarising product reviews is one of the most researched text-based UGC summarisation topics. Sentiment-based quantitative summaries, e.g. percentage positive vs negative or star ratings are covered in Section 14.5. With respect to text-based review summaries, there has been work on aspect-based summarisation [88, 15], comparative summaries of contradictory opinions, and ultra-concise summaries.

In terms of approaches, abstractive summarisation is much more common than extractive approaches, largely due to the specifics of product reviews. In particular, Carenini *et al* [15] have argued that the abstractive summarisation paradigm achieves better results on product reviews than extractive summarisation techniques.

One kind of product review summarisation that has been studied is *aspect-based summaries*. This kind of summaries consist of a number of aspects or product features (e.g. food, price for a restaurant), a numeric value for the overall aspect rating from the reviews, and a set of textual snippets highlighting opinions on each of the aspects (e.g. 'great price', 'awful waiter'). As discussed in [88], aspect-based summarisation needs to solve two problems: *aspect mention extraction* (e.g. service from the phrase 'awful waiter') and *sentiment classification* (e.g. awful as being negative). Sentiment is also aggregated for each aspect, in order to produce a numeric value for the overall sentiment across all product reviews (e.g. service 1 star). For example, [41] use association mining to identify frequently mentioned product features, coupled with using opinion words as context for identifying infrequent ones. More recently, [88] studied the problem of identifying all mentions of product aspects (i.e. features) at a phrase level and using these textual snippets in the aspect summaries. This abstractive summarisation method, called multi-aspect sentiment model, assigns words from the reviews to aspects using a topic model, coupled with aspect-specific maximum entropy classifiers which predict the sentiment rating towards each aspect. Carenini *et al* [15] instead propose an approach based on Natural Language Generation (NLG) techniques, which produce more qualitative aspect summaries of opinions (e.g. 'Customers had mixed opinions about the Apex AD2600..'). In addition, they define a *controversiality* measure for a set of opinions and demonstrate that NLG-based abstractive summaries are better than extractive ones, especially on highly controversial topics.

Another UGC summarisation problem is that of *contrastive opinion sum-*

marisation, where the aim is to extract contrastive pairs of sentences, which contain contradictory opinions. [46] argue that such summaries are helpful for people wanting to digest mixed product reviews, e.g. some reviews may rate iPhone battery life highly, while others might criticise it. Kim and Zhai [46] formulate the problem as an extractive summarisation task, where positive and negative sentences are chosen based on two criteria. The first is representativeness, i.e. a similar positive/negative opinion must be expressed in many reviews, and the second is contrastiveness, i.e. the contradictory opinions must be on the same aspect or feature of the product. More recently, Paul *et al* [71] have investigated summarisation of contrastive viewpoints of political opinions, e.g. points raised by people for and against a new healthcare reform. The algorithm has two stages. First it clusters opinionated texts by viewpoint based on a topic model – a step called macro viewpoint summarisation. Then a micro viewpoint summary is generated, containing multiple sets of contrastive sentences.

Going below the sentence level, researchers have studied the problem of generating *very concise summaries*. [33] present an unsupervised approach, which given a product review (tested on reviews longer than 5 sentences) selects a single *supporting sentence*, which best captures the overall sentiment of that review. Another kind is the pros and cons product summary, which aims to convey the sentiment expressed by the majority of product reviews. Branavan *et al* [12] develop a model that assigns keyphrases to product reviews and generates a pros and cons list of these phrases, which tend to be indicative of product properties. The Opinosis system [30] goes one step further by generating text-based, pros and cons summaries, using a graph-based method for abstractive summarisation. The method utilises the highly redundant nature of product reviews, in order to extract the most representative phrases. The most recent abstractive approach [31] focuses on micropinion review summaries, which in addition to product properties, also include sentiment-bearing adjectives (e.g. very short battery life, big screen). Representative phrases are chosen using point-wise mutual information and scored for readability using the Microsoft n-gram service.[2] The method improves on the results of Opinosis and is evaluated on user-generated reviews for 330 products from CNET. The ROUGE evaluation metric [50] is used for quantitative evaluation, while human assessors also assigned qualitative scores for grammaticality, non-redundancy, and informativeness. Although the method has only been evaluated on product reviews, it should generalise to summarising other types of short UGC, e.g. tweets, comments.

14.4.2 Blog Summarisation

Unlike product reviews, blog post summarisation has been approached mostly through extractive summarisation techniques and modelled as the

[2] http://web-ngram.research.microsoft.com

problem of selecting the most representative sentences from one or more blog posts (i.e. single-document vs multi-document summarisation).

In order to determine which sentences need to be kept in the blog summary, Zhou and Hovy [98] made use of the content of news articles, linked from the blog post. In other words, only sentences in the blog post, which are similar to the linked articles were retained. This approach was, however, only tried on political blogs and it remains unclear whether it would generalise to other kinds of links to external web content. The authors raise the question of evaluating the quality of blog post summaries and propose, but do not carry out, extrinsic, task-based evaluation instead.

One of the first papers on *opinion-based blog summarisation* is by Ku *et al* [47], who distinguish between negative and positive documents for a given topic, based on the topics and sentiment expressed in the individual sentences. A brief positive/negative summary of these two sets of documents is generated based on the document headlines. The detailed positive/negative summaries consist of representative sentences with the required polarity, i.e. an extractive summarisation approach.

Further research on this topic has been driven by the TAC 2008 opinion summarisation task[3], which built a gold standard dataset of summaries of opinions expressed in a given set of blog posts, about a given target (i.e. a multi-document summarisation problem). For instance, Schilder *et al* [82] approached the TAC 2008 opinion-based blog summarisation task by extending the FastSum extractive multi-document summarisation system with a lexicon-based sentiment tagging module and blog-specific pre-processing. In a follow-up work [21], FastSum is applied to legal blogs and made less dependent on the TAC 2008 task specifics. In more detail, following the blog pre-processing step, the method filters out sentences from the blog posts, which do not match the given opinion target (a named entity in TAC 2008 and a noun phrase in the legal blogs). The remaining sentences are sentiment tagged, based on a lexicon of positive and negative words. An SVM classifier is trained to rank the sentences for inclusion in the opinion summaries, based on various word frequency features, as well as sentence length and position features.

Using the TAC 2008 corpus, [61] carry out an error analysis of opinion-based blog summarisation in comparison to news summarisation and note that the latter is an easier task, where automated methods score higher. This is attributed to the differences between the two genres, with blogs containing much more pronounced opinions, being noisier, and not having a stereotypical structure (e.g. sentence position in news articles is a very important feature). A finer-grained analysis showed that errors in the summaries related to topic irrelevancy, incoherent discourse, presence of irrelevant information and syntactic and lexical mistakes are much more frequent in blogs than in news.

[42] address the problem of *comments-oriented blog summarisation*, i.e. that of selecting representative sentences from a blog post, by leveraging the

[3] http://www.nist.gov/tac/2008/summarization/op.summ.08.guidelines.html

topics covered in the associated comments. Sentence selection is based on the representativeness scores of the contained words, normalised for sentence length. Comment-based word representativeness scores are computed in four ways: binary (whether a word appears in a comment); comment frequency (number of comments containing the word); term frequency; and the best performing ReQuT (Reader, Quotation, and Topic) model. The reader measure is based on how often a user mentions other users in the comments; quotation captures comments quoting other comments; whereas comments topics are discovered via clustering and scored for importance via cosine similarity to the cluster centroid. Even though the ReQuT model was defined originally for blog summarisation, it could be applied also to the problem of summarising a set of tweets, due to the fact that tweets have similar user mention networks, can quote or reply to other tweets, and can be clustered around common topics (e.g. [78]).

[85] addressed the problem of *summarising the blog network* by finding the most influential blogs and the opinions they contain. Given a search query (e.g. YouTube), a network of relevant blogs is constructed, where nodes correspond to blogs and edges correspond to links between blog posts in the respective blogs. The method ranks blogs for importance to other blogs and the novelty of the information contained in the posts. Hassan *et al* [40] investigate a different graph based approach which uses lexical centrality and is based on the LexRank [27] extractive summarisation algorithm.

14.4.3 Summarising Very Short UGC

As discussed in Section 14.3, summarising tweets and other kinds of short UGC (e.g. comments on a news or a video site) is a particularly challenging problem. Here we will discuss methods that cast short message summarisation as an extractive summarisation problem, where the goal is to select the most representative subset of posts. Other ways to summarise short UGC are to produce high-level overviews based on topics (see Section 14.6) or sentiment (see Section 14.5).

Firstly, looking at Twitter, Inouye *et al* [44] have compared several summarisation algorithms on producing multi-post summaries of tweets, where the four most informative tweets are selected as the summary. A Hybrid TF.IDF algorithm is defined, where term frequencies are computed by regarding all tweets as a single document, but IDF computation considers each tweet as a separate document. This is motivated by the shortness of tweets (and status updates in general), which makes it very unlikely that a word appears more than once in a tweet. The authors collect tweets containing a trending hashtag and observe that even though there is a common overall topic, the posts often tend to cluster around several aspects or sub-topics. Therefore, they also propose a clustering-based approach, which first applies k-means clustering (with k set to 4) and then for each cluster of tweets, their Hybrid TF.IDF algorithm selects the best tweet. Inouye *et al* [44] compared a number of summarisation

algorithms, including a random baseline; a baseline selecting the most recent tweets; the MEAD multi-document summariser [76]; LexRank [27]; and their own cluster-based and Hybrid TF.IDF algorithms. The results (both quantitative and human-produced scores) showed that the frequency-based extractive summarisation algorithms, such as their Hybrid TF.IDF, did best, whereas centroid and graph-based algorithms did not perform well, most likely due to tweets being standalone and short.

A similar problem is summarising user comments on news sites, YouTube, and other social web sites. [45], in particular, studied extractive comment summarisation on YouTube and formulate the problem as selecting the most representative top k comments, i.e. the same as [44] do for tweet summaries. [45] also first cluster the user comments, then apply precedence-based ranking to select the most representative comments for each of the clusters. Two clustering approaches were tried – k-means (used also by [44] above) and LDA-based topic clustering – with the latter producing better results. Different comment selection methods were tried: TF.IDF, mutual information, MEAD [76] and LexRank [27]. Mutual information outperformed TF.IDF and again, MEAD and LexRank did not perform as well.

One particular kind of short message (or microblog) summarisation is event-based summarisation, where the structure of real-world events can be used to help with the selection and ordering of the most relevant microblog content. For summarisation of such short UGC sets, [8] select tweets based on their textual quality (e.g. no spelling mistakes), relevance to the event, and usefulness for conveying details about the event. The top 5 short posts are selected on that basis, using methods based on centroid similarity, degree centrality and LexRank [27]. Their findings indicate that the centroid method is significantly better than the other two on the relevance and usefulness criteria, with no significant difference between degree centrality and LexRank. Again, similar to the methods discussed above, only tweet content is used as input, in a bag of words fashion.

However, as discussed in Section 14.3, microblogs and social network updates in particular, contain a wealth of implicit semantics which could be used to improve the performance of purely text-based summarisation methods. In particular, information propagation metadata in tweets can be used as additional source of relevance and importance. More specifically, Harabagiu and Hickl [38] focus on retweets, responses (a user responds to another user's post), and quote chains (where a post quotes from another). The method also makes use of richer semantic knowledge from the microposts themselves, namely named entities, event mentions, temporal information, and inter-event relationships (identity and temporal precedence).

In addition to exploring the graph-like relationships between tweets (e.g. retweets, replies, mentions), researchers have also started to exploit information from the social user networks. For instance, [93] propose a graph co-ranking method which makes use of the following relations in Twitter to form a user network graph; a micropost network based on content similarity and

relationships between posts; and a bi-partite graph which connects the two kinds of networks. The co-ranking approach is motivated by the connection between the content of a post, who posted it, and what is their standing in the social network (e.g. celebrities have millions of followers on Twitter).

Returning to the problem of event-based summarisation, researchers have also studied ways to generate useful thematic/topical descriptors for automatically discovered sub-events. For example, in the context of sub-events of football games (e.g. red card, goal), [99] select one most representative tweet per sub-event. They compare a term frequency count in tweets occurring in the minute of the sub-event against the Kullback-Leibler divergence metric, which not only captures frequency within a sub-event, but also takes into account the overall frequency in the entire event-related set of tweets, up until the new sub-event occurred. The latter method was shown to produce consistently better results, also evaluated across 3 languages: English, Spanish, and Portuguese. Others [**?**] have used point-wise mutual information, coupled with user geolocation and temporal information, in order to derive n-gram event descriptors from tweets. By making the algorithm sensitive to the originating location, it is possible to see what people from a given location are saying about an event (e.g. those in the US), as well as how this differs from tweets elsewhere (e.g. those from India). Similarly, the temporal information results in different text descriptors being extracted on different days, as the event unfolds.

14.5 Structured, Sentiment-Based Summarisation of UGC

The focus of this section is on methods for generating structured summaries, based on quantitatively aggregated sentiment from UGC. Naturally, the first step of such methods is *sentiment detection*, followed by an *quantitative summarisation* step.

Recently, techniques for sentiment detection and aggregation have begun to focus on UGC, combined with a trend towards its application as a proactive rather than a reactive mechanism. Understanding public opinion can have important consequences for the prediction of future events.

It is beyond the scope of this chapter to provide an in-depth review of automatic sentiment detection techniques, instead we refer the reader to [68]. In general, sentiment detection techniques can be roughly divided into *lexicon-based methods*, e.g. [75, 81, 87] and machine-learning methods, e.g. [69, 89, 10, 67, 34]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine learning approaches make use of syntactic and/or linguistic features [67, 34], and hybrid approaches are

very common, with sentiment lexicons playing a key role in the majority of methods, e.g. [22].

With respect to quantitative aggregation of sentiment and opinions, especially coupled with monitoring over time, methods can vary greatly in their degree of sophistication. The simplest approach is to produce overall statistics of positive vs negative opinion, e.g. in a given set of documents or towards a given entity, such as a famous politician [48] or a product [41]. A slightly more complex approach is to calculate quantitative statistics within specific time intervals, e.g. daily, weekly. In some applications, more detailed approaches are needed, such as modelling the opinion holders and strength of conflicting opinions and how they change over time.

A number of methods are surveyed next, grouped according to the type of user-generated content they focus on.

Starting with product reviews, [41] propose a semi-structured, aspect-based summarisation approach to quantitative sentiment summarisation. For each sentence in a product review the method discovers which feature is talked about, then categorises the sentences into positive vs negative, and increments the aggregated sentiment for this feature. The result, for example, on a set of digital camera reviews would be: picture quality (positive 253; negative 6); size (positive 134; negative 10).

Sentiment detection and aggregation research has also gone beyond the basic positive vs negative sentiment, towards assigning multi-point ratings to the entire review [70] or at aspect level (called *rated aspect summaries* [51]), e.g. [51, 36, 52]. These ratings can be viewed as stars (e.g. 1 to 5 stars) or discrete values and aggregated across multiple reviews for the same product or shown on a review by review basis.

Similar methods have also been studied on Twitter data. [67] classify arbitrary tweets on the basis of positive, negative and neutral sentiment and then aggregate the overall sentiment. They construct a binary classifier which used n-gram and POS features, and train it on instances, which had been annotated according to the existence of positive (':)') and negative (':(') emoticons. Their approach has a lot in common with an earlier sentiment classifier constructed by [34], which also used unigrams, bigrams and POS tags, though the former demonstrated through analysis that the distribution of certain POS tags varies between positive and negative posts.

[48] tackles a somewhat different sentiment analysis and aggregation task. Tweets relating to president Obama are analysed and a daily overall "strong sentiment" is calculated. This figure is given as the ratio of the count strongly positive tweets over the strongly negative ones. The strength and polarity of Tweets in the dataset is calculated according to learned lexicons, which are lists of keywords which in general correspond to either positive or negative sentiment.

[22] also made use of a sentiment lexicon to annotate and aggregate positive and negative sentiment in tweets related to political events. They performed supervised learning with manually annotated examples to train a binary clas-

sifier of political opinion, using this second classifier when the former failed to make a classification. They only report the overall sentiment from a collection of Tweets during a specific time-window, and their system will refrain from reporting sentiment when no concensus appears to be reached for that period.

There also exists a plethora of commercial search-based tools for performing quantitatively aggregated sentiment analysis of tweets. Generally, the user enters a search term and gets back all the positive and negative (and sometimes neutral) tweets that contain the term, along with some graphics such as pie charts or graphs. Typical basic tools are Twitter Sentiment[4], Twends[5] and Twitrratr[6]. Slightly more sophisticated tools such as SocialMention[7] allow search in a variety of social networks and produce other statistics such as percentages of Strength, Passion and Reach, while others allow the user to correct erroneous analyses. On the surface, many of these appear quite impressive, and have the advantage of being simple to use and providing attractive quantitative summaries with copious information about trends. However, such tools mostly aim at finding public opinion about famous people, sports events, products, movies and so on, but do not lend themselves easily to more complex kinds of opinion or to more abstract kinds of searches. Furthermore, their analysis tends to be fairly rudimentary, performance can be quite low, and many of them do not reveal the sources of their information or enable any kind of evaluation of their success: if they claim that 75% of tweets about Whitney Houston are positive, or that people on Facebook overwhelmingly believe that Greece should exit the eurozone, we have no proof as to how accurate this really is.

14.6 Keyword-based Summarisation of UGC

While textual summarisation typically attempts to address not only the content of a summary but also its properties as a single piece of text, such as its coherence and cohesiveness, there is a specific formulation of the task of summarisation in which these issues are avoided altogether. In keyword extraction, a summary is simply a list of terms which can be considered salient.

Automatically selected keywords are useful in representing the topic of a document or collection of documents, and less effective in delivering arguments or full statements contained therein. It is therefore necessary to consider keyword extraction as a form of indicative summarisation, allowing the reader to decide whether or not to view the full text. Keywords can also be used in the context of information retrieval, as a means of dimensionality reduction

[4] http://twittersentiment.appspot.com/
[5] http://twendz.waggeneredstrom.com/
[6] http://twitrratr.com/
[7] http://socialmention.com/

and allowing systems to deal with smaller sets of important terms rather than whole documents.

In early summarisation, terms with high frequency within a document were assumed to be important and sentences which contained these terms were favoured for inclusion in a summary [53]. Additionally, [13] used TF.IDF to characterise important terms in a document. Both these approaches share the property that they rely on models of term significance.

Later approaches to keyword extraction exploited term co-occurrence; forming a graph of terms with edges derived from the distance between occurrences of a pair of terms and assigning weights to vertices [60]. This class of keyword extraction was found to perform favourably on Twitter data compared to methods which relied on text models [91].

These graph-based approaches to extracting keywords from Twitter perhaps perform well because the domain contains a great deal of redundancy [96]. While this property of Twitter is somewhat beneficial when producing keyword summaries, another less helpful trait is the sheer variety of topics discussed on the service. When it is not known that a document discusses a single topic, it can be more difficult to extract a coherent and faithful set of keywords from it.

Personal twitter timelines, when treated as single documents, present this problem. Users are generally capable of posting on multiple topics. While [91] use TextRank on the whole of a user's stream, they do not attempt to model or address topic variation, unlike [92], who incorporated topic modelling into their approach. Theirs is not the only application of Topic Modelling to Twitter data, as it is similar to [78]. However in the latter work topics are discovered, but never summarised.

Document clusters can still be somewhat difficult for humans to read, containing a great many tweets. For summarisation of these document sets, [8] evaluate methods based on centroid, degree centrality and LexRank, finding that centroid works the best in some cases, but without statistically significant differences.

Larger threads, like trending topics on Twitter, tend to contain a great deal of redundancy thanks to retweeting of messages and copying-and-pasting. [84] extracted keyphrases for trending topics by exploiting textual redundancy and selecting common sequences of words. Their short phrases are similar to the pithy, manually generated summaries created by users of services like WhatTheTrend.

It could be argued that many events, once successfully extracted from Twitter, remain too coarse to be used for informative summaries. Though [8] investigate the effectiveness of various traditional summarisation methods when applied to clusters of tweets belonging to events, the granularity of an event is somewhat unpredictable; while the TDT initiative defines an event as "Something interesting which occurs at a specific time and place" [2], in practise this constraint is difficult to enforce.

14.7 Visual Summarisation of UGC

Twitter data can be summarised visually as well as textually. These 'visual summaries' can take many forms, including graphs and charts, maps and timelines. In this section we will briefly discuss the forms which non-textual summaries of UGC can take in existing systems.

One of the simplest and widely used summary visualisations is word clouds. These generally use single word terms, which can be somewhat difficult to interpret without extra context. Word clouds have been used to assist users in browsing social media streams, including blog content [6] and tweets [83, 63]. For instance, Phelan *et al* [73] use word clouds to present the results of a Twitter based recommendation system. The Eddi system [9] uses topic clouds, showing higher-level themes in the user's tweet stream. These are combined with topic lists, which show who tweeted on which topic, as well as a set of interesting tweets for the highest ranked topics. The Twitris system (see Figure 14.3) derives even more detailed, contextualised phrases, by using 3-grams, instead of uni-grams [63].

The main drawback of cloud-based visualisations is their static nature. Therefore, they are often combined with timelines showing keyword/topic frequencies over time [1, 9, 43, 90], as well as methods for discovery of unusual popularity bursts [6]. [22] use a timeline which is synchronised with a transcript of a political broadcast, allowing navigation to key points in a video of the event, and displaying tweets from that time period. Overall sentiment is shown on a timeline at each point in the video, using simple colour segments. Similarly, TwitInfo (see Figure 14.1 [57]) uses a timeline to display tweet activity during a real-world event (e.g. a football game), coupled with some example tweets, colour-coded for sentiment. Some of these visualisations are dynamic, i.e. update as new content comes in (e.g. topic streams [23], falling keyword bars [43] and dynamic information landscapes [43]).

In addition, some visualisations try to capture the semantic relatedness between topics in the media streams. For instance, BlogScope [6] calculates keyword correlations, by approximating mutual information for a pair of keywords using a random sample of documents. Another example is the information landscape visualisation, which convey topic similarity through spatial proximity [43] (see Figure 14.2). Topic-document relationships can be shown also through force-directed, graph-based visualisations [24]. Lastly, Archambault *et al* [3] propose multi-level tag clouds, in order to capture hierarchical relations.

Another important dimension of user-generated content is its place of origin. For instance, some tweets are geo-tagged with latitude/longitude information, while many user profiles on Facebook, Twitter, and blogs specify a user location. Consequently, map-based visualisations of topics have also been explored [59, 57, 43, 63] (see also Figures 14.2 and 14.1). For instance, Twitris

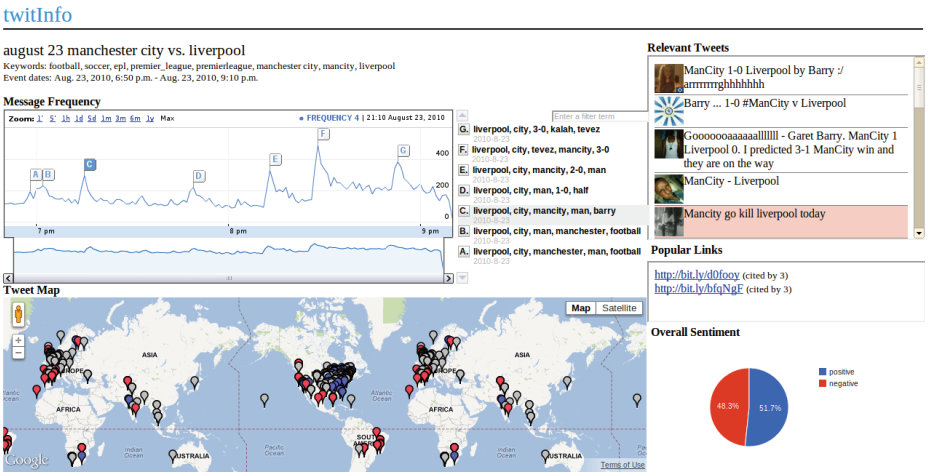

FIGURE 14.1: TwitInfo tracks a football game (http://twitinfo.csail.mit.edu/)

[63] allows users to select a particular state from the Google map and it shows the topics discussed in social media from this state only. Figure 14.3 shows the Twitris US 2012 Presidential elections monitor, where we have chosen to see the related topics discussed in social media originating from California. Clicking on the topic "female democratic senators" displays the relevant tweets, news, and Wikipedia articles.

Opinions and sentiment also feature frequently in UGC summarisation interfaces. For instance, Media Watch (Figure 14.2 [43]) combines word clouds with aggregated sentiment polarity, where each word is coloured in a shade of red (predominantly negative sentiment), green (predominantly positive), or black (neutral/no sentiment). Search results snippets and faceted browsing terms are also sentiment coloured. Others have combined sentiment-based colour coding with event timelines [1], lists of tweets (Figure 14.1 [57]), and mood maps [1]. Aggregated sentiment is typically presented using pie charts [90] and, in the case of TwitInfo, the overall statistics are normalised for recall (Figure 14.1 [57]).

Researchers have also investigated specifically the problem of summarising visually social media conversations about real-world events, e.g. broadcast events [83], football games (Figure 14.1 [57]), conferences [23], and news events [59, 1]. A key element here is the ability to identify sub-events and combine these with timelines, maps, and topic-based visualisations.

Lastly, given the user-generated and social nature of the media streams, some visualisations have been designed to exploit this information. For instance, the PeopleSpiral visualisation [23] plots Twitter users who have contributed to a topic (e.g. posted using a given hashtag) on a spiral, starting with

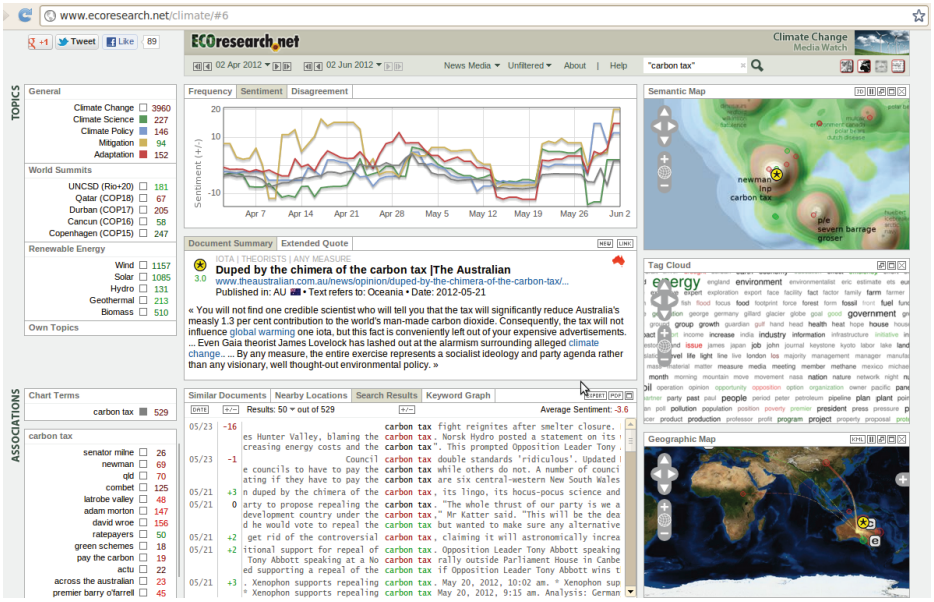

FIGURE 14.2: Media Watch on Climate Change Portal (http://www.ecoresearch.net/climate)

the most active and 'original' users first. User originality is measured as the ratio between the number of tweets authored by the user versus re-tweets made. OpinionSpace [28] instead clusters and visualizes users in a two-dimensional space, based on the opinions they have expressed on a given set of topics. Each point in the visualisation shows a user and their comment, so the closer two points – the more similar the users and opinions are. However, the purely point-based visualisation was found hard to interpret by some users, since they could not see the textual content until they clicked on a point. ThemeCrowds [3] instead derives hierarchical clusters of Twitter users through agglomerative clustering and provides a summary of the tweets, generated by this user cluster, through multilevel tag clouds (inspired by treemap visualisation). Tweet volumes over time are shown in a timeline-like view, which also allows the selection of a time period.

14.8 Evaluating UGC summaries

In any domain, the task of summarisation can be a difficult one to evaluate. Summaries can take many different forms, and even for the common case of

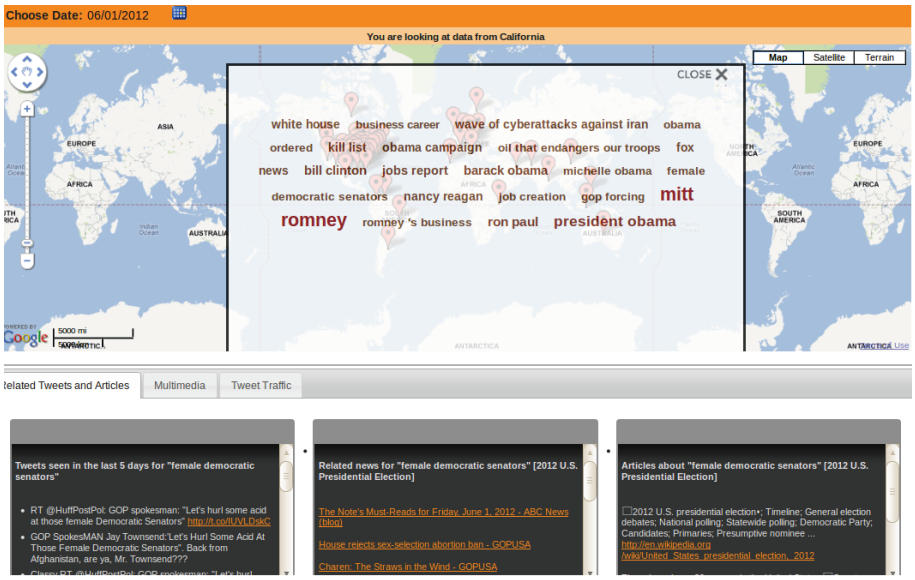

FIGURE 14.3: The Twitris Social Media Event Monitoring Portal (http://twitris.knoesis.org)

textual summarisation it is difficult to automatically compare summaries to a gold standard. Issues arise from the ability to validly represent a compressed form of a document in any of a number of ways. The problem can be even more acute on short UGC, such as tweets, where summaries may take the form of lists, or visual representations of tweet volume or other types of timeline.

Additionally, since it is generally impossible to read all of the short messages for a corpus to be summarised, gold-standard human generated summaries can be very incomplete or may miss information which should have been included.

The difficulty of evaluating visual summaries of UGC is so acute that many choose not to address it at all [90, 83, 63, 20]. Visual summaries typically contain information other than raw text, for example they may display information from tweets on a map, or include timelines of topics and tweet volume. They may include less local information, like news stories or video. The variability of such summary makes them very difficult to evaluate using a single standard method. Addressing this very issue, [22] carry out a user study of their system, performing extrinsic testing by having journalists attempt to use the summariser to form ideas about possible articles. The 18 journalists taking part were asked to complete surveys about the experience, and the stories they produced were assigned individual types. Though the authors admit that evaluation on such a small set of users is not enough to truly test the efficacy of their summariser, they argue that the experiment has allowed them

to see the ways in which it could be used, and they believe that testing on a larger set of users could lead to more significant results.

For truly textual UGC summaries on relatively small test corpora, authors can reasonably hope to use the standard DUC evaluation measures on their systems. For instance, [84] gather 100 tweets per topic for 50 trending topics on Twitter, and then ask volunteers to generate summaries which they feel best describe each collection. The automatic summaries produced by their algorithm are then tested using a manual content annotation method, testing that the same content is available as in the gold standard, and automatically using Rouge-1. [39] again make use of model summaries, though they eschew Rouge in favour of the Pyramid method [66].

Another class of evaluation treats summarisation as a document retrieval problem. [17] use knowledge about the significant plays in a sports match to build a gold standard for what must be discussed in a 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.

[8] similarly have users score the quality, relevance and usefulness of the selected tweet summary, though they do not attempt to address the problem of calculating recall.

14.9 Outstanding Challenges

Within the scope of this chapter, we have discussed the current strands of research into summarising user-generated content. Many of the methods discussed here have not traditionally been considered types of summarisation, e.g. the creation of keyword-based and visual summaries of UGC.

We believe that summarisation for Twitter and similar online social networks is a problem of the utmost practical relevance. While users post an enormous wealth of extremely useful UGC with frightening regularity, it is simply impossible to filter and manually process all but the smallest data stream. Where the task of summarisation might once have been to find the salient points in a document collection, it is now required to find the salient points to a particular user, on a particular topic, in the entire of the Twitterverse.

This need for compression and summarisation - the requirement for users to be shown what is relevant to them and little else, has driven our interpretation of many existing strands of work in the light of UGC summarisation. In the remainder of this chapter we discuss several major outstanding research challenges.

14.9.1 Spatio-Temporal Summaries

Even though UGC is strongly grounded in the spatio-temporal context, this additional semantic information has so far mostly been neglected in UGC summarisation. For instance, where visualisation is concerned, it is mostly limited to map-based and topic-based timeline visualisations (see Section 14.7). Additional summarisation dimensions can be users' age, gender, political views, interests, and other such latent characteristics (see Section 14.9.4 below). In addition, UGC summarisation methods based on social networks (e.g. hubs and authorities) could be combined with the currently prevailing topic- and content-centric approaches.

The main sources of user location information are GPS coordinates attached to a message and self-disclosed location information in user profiles. However, [19] found that only around 36% of users actually provide a valid location in their profiles. Furthermore, when we analysed a dataset of over 30,000 tweets discussing the 2011 London Riots, less than 1% of microposts contained any GPS information.

Therefore, an important pre-requisite for location-based summarisation is the automatic geolocation of users of social networks, based on publicly available data, e.g. their profile, posts and social network. Broadly speaking, existing methods fall into two different categories: content-based (i.e. using the textual posts of a given user) and network-based (i.e. using the social network).

Content-based methods ('you are where you write about') typically gather the textual content produced by the user and infer their location based on features, such as mentions of local place names [29] and use of local dialect. In the work of [25, 19], region-specific terms and language that might be relevant to the geolocation of users were discovered automatically. A classification approach is devised in [55] that also incorporates specific mentions of places near to the user. One obvious disadvantage to this method is the fact that someone might be writing about a popular global event which is of no relevance to his actual location. Another is that users might take deliberate steps to hide their true location by alternating the style of their posts or not referencing local landmarks.

In contrast, network-based geolocation methods ('you are where your friends are') aim to use the user's social network to infer their location. To the best of our knowledge, the only existing method of this kind (i.e., relying on the user's social network alone) is the work of [4], who first create a model for the distribution of distance between pairs of friends, before using this to find the most likely location for a given user. The influence of distance in social network ties is demonstrated by the earlier work of [49]. Two limitations of this approach are that it assumes all users globally have the same distribution of friends in terms of distance and fails to account for the density of people in an area. In conclusion, automatic user geolocation is still a relatively under-

explored problem, but nevertheless, where known, location information could usefully be incorporated in UGC summaries.

One example application, needing spatio-temporal summaries, is monitoring political elections and the popularity of political parties and politicians over time. These not only vary on a local, regional, and national level, but also change over time. As part of the TrendMiner[8] project, we are currently developing methods for spatio-temporal summarisation of Twitter streams.

14.9.2 Exploiting Implicit UGC Semantics

Although inroads have been made already, current methods for UGC summarisation have many limitations. Firstly, most methods address the more shallow problems of keyword and topic extraction, while frequency-based extractive summarisation techniques do not reach the significantly better performance obtained by more sophisticated methods on longer text documents.

Secondly, the majority make very few or no adaptations to better model the specifics of user-generated content. UGC abounds with implicit semantic information, which is very different from the discourse and positional features in traditional text documents. More specifically, UGC summarisation methods need to tap better into the knowledge from the user social networks (who is connected to who), information diffusion networks (who retweeted, replied, and mentioned who), the user's own produced UGC, temporal information (e.g. recency), and user demographics. For instance, one could envisage a new kind of opinion-based summaries, which summarise opinions, according to influential groups, demographics and geographical and social cliques.

14.9.3 Multilinguality

With fewer than 50% of tweets in English [16], another related major challenge is multilinguality. Most of the methods surveyed here were developed and tested on English content, with some exceptions, e.g. [99] who also considered Spanish and Portuguese for event-based summarisation.

Multilingual summarisation is a challenge even in the context of standard, well-formed documents. As recently as 2011 the Text Analysis Conference (TAC) ran the MultiLing pilot competition[9], specifically with the aim to promote the development of multilingual algorithms for summarisation. The task was to create short summaries of around 240 words of 10 news texts on a given topic. Seven languages were covered and each system had to produce summaries in at least two languages. The best performing system [86] used a summarisation algorithm based on latent semantic analysis, in order to determine the highest scoring sentences for the summaries. The only language-specific resources were stop word lists. The question of how well such an

[8] http://www.trendminer-project.eu
[9] http://users.iit.demokritos.gr/ ggianna/TAC2011/MultiLing2011.html

approach would cope with user-generated content and the different kinds of UGC summaries that could be produced, requires further experimentation.

14.9.4 Personalisation

The forth major challenge is the generation of personalised summaries of UGC. As discussed in Section 14.9.1 above, the user's profile and the content they contribute online to the various social networking sites can be a useful source for deriving information about the user and using that to personalise the summaries.

For instance, [93] have recently proposed a way to inject a topic-based model of user interests in their tweet recommendation system. There have also been efforts to discover user demographics information, when it is not available already. [14] classify users as male or female based on the text of their tweets, their description fields and their names. They report better-than-human accuracy, compared to a set of annotators on Mechanical Turk. [72] present a general framework for user classification which can learn to automatically discover political alignment, ethnicity and fans of a particular business.

With respect to capturing user interests from tweets, further work is required on distinguishing globally interesting topics (e.g. trending news) from interests specific to the given user (e.g. work-related, hobby, gossip from a friend, etc.).

What is interesting to a user also ties in with user behaviour roles. In the case of online forums, the following user behaviour roles have been identified [18]: *elitist*, *grunt*, *joining conversationalist*, *popular initiator*, *popular participant*, *supporter*, *taciturn*, and *ignored*. In Twitter, the most common role distinction is between *meformers* (80% of users) and *informers* (20% of users) [62].

In turn, this requires more sophisticated methods for the automatic assignment of user roles, based on the semantics of posts and user interaction patterns, as well as the successful integration of these features into the UGC summarisation algorithms.

14.9.5 Training Data, Evaluation and Crowdsourcing

Two of the major stumbling blocks faced by researchers working on UGC summarisation methods are the lack of lack of training data and the need for human assessors for evaluating the quality of the generated summaries. Consequently, crowdsourcing, primarily using Amazon Mechanical Turk, is being tried as means to address these two bottlenecks.

When humans are asked to create a text summary for one or more given documents, this is effectively a content generation problem. Gathering these through crowdsourcing has been shown to be particularly challenging, since naive task definitions produce low-quality corpora [95, 64].

Consequently, the design and quality control of the summary collection crowdsourcing tasks is non-trivial. This is only partly due to cheating, but mainly due to the fact that the crowdworkers generate multiple diverse answers, all of which could be correct, i.e. different, good summaries can be provided for the same UGC content. For instance, [26] create a corpus of extractive summaries for Arabic texts collecting three summaries per text where summaries consist of a sub-set of representative sentences from the article. They use majority voting to determine which sentences need to be kept in the summary, i.e. at least two of the workers need to agree.

When crowdsourcing is used instead to crowdsource human evaluations of summary quality, researchers are reporting mixed experiences so far.

For instance, [32] have found that non-expert evaluation of summarisation systems produces noisier results thus requiring more redundancy to achieve statistical significance and that MTurk workers cannot produce score rankings that agree with expert ranking. The authors suggest that redesigning the AMT task definition according to other summarization evaluation approaches could improve results. Indeed, a key challenge in using MTurk for summary evaluation is to transform a complex expert-based evaluation protocol (which typically relies on detailed instructions) into smaller, simpler tasks that can be explained to non-experts [58].

One simpler task design has been used by [44] for evaluation of tweet summaries. Inouye *et al* asked the MTurk workers to indicate on a five point scale, how much of the information from the human produced summary is contained in the automatically produced summary. Another simpler task design that has achieved successful results on MTurk is pair-wise ranking [33]. The summarisation task in this case is to identify the most informative sentence from a product review. In this case, the crowdworkers are asked to indicate whether a sentence chosen by the baseline system is more informative than a sentence chosen by the author's method. Sentence order is randomised and it is also possible to indicate that none of these sentences are a good summary.

Overall, even though crowdsourcing is increasingly playing a role in summarisation research, using it successfully still requires significant expertise. Further research is required in making the process scalable, repeatable, and capable of producing high quality summaries and evaluation results.

The next step forward would be to make freely available reusable task definitions and crowdsourcing workflow patterns, as well as providing more details in the research papers on the exact methodology that was followed.

Acknowledgements: This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1).

Bibliography

[1] B. Adams, D. Phung, and S. Venkatesh. Eventscapes: visualizing events over time with emotive facets. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1477–1480, 2011.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45, 1998.

[3] D. Archambault, D. Greene, P. Cunningham, and N. J. Hurley. Themecrowds: multiresolution summaries of twitter usage. In *Workshop on Search and Mining User-Generated Contents (SMUC)*, pages 77–84, 2011.

[4] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, pages 61–70. ACM, 2010.

[5] T. Baldwin and M. Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California, June 2010.

[6] N. Bansal and N. Koudas. Blogscope: Spatio-temporal analysis of the blogosphere. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1269–1270, 2007.

[7] Regina Barzilay and Kathleen R. Mckeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31:297–328, 2005.

[8] H. Becker, M. Naaman, and L. Gravano. Selecting Quality Twitter Content for Events. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[9] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd ACM Symposium on User Interface Software and Technology (UIST)*, pages 303–312, 2010.

[10] E. Boiy and M-F. Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558, 2009.

[11] K. Bontcheva and Y. Wilks. Tailoring Automatically Generated Hypertext. *User Modeling and User-Adapted Interaction*, 2004. Special issue on Language-Based Interaction.

[12] S. R. K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, 34(1):569–603, April 2009.

[13] Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, September 1995.

[14] J. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, 2011.

[15] Giuseppe Carenini and Jackie Chi Kit Cheung. Extractive vs. nlg-based abstractive summarization of evaluative text: the effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 33–41, 2008.

[16] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, Forthcoming.

[17] D. Chakrabarti and K. Punera. Event Summarization Using Tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[18] J. Chan, C. Hayes, and E. Daly. Decomposing discussion forums using common user roles. In *Proceedings of WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[19] Z. Cheng. You are where you tweet: A content-based approach to geolocating twitter users. *Proceedings of the 19th ACM Conference*, 2010.

[20] Brendan O. Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the Fourth AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 384–385, 2010.

[21] Jack G. Conrad, Jochen L. Leidner, Frank Schilder, and Ravi Kondadadi. Query-based opinion summarization for legal blog entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 167–176, 2009.

[22] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 115–122, 2010.

[23] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, November 2010.

[24] J. Eisenstein, D. H. P. Chau, A. Kittur, and E. Xing. Topicviz: Semantic navigation of document collections. In *CHI Work-in-Progress Paper (Supplemental Proceedings)*, 2012.

[25] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[26] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Using mechanical turk to create a corpus of arabic summaries. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010.

[27] Günes Erkan and Dragomir R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal Artificial Intelligence Research*, 22(1):457–479, 2004.

[28] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI)*, pages 1175–1184, 2010.

[29] Clay Fink, Christine Piatko, James Mayfield, Tim Finin, and Justin Martineau. Geolocating blogs from their textual content. In *Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 1–2. AAAI Press, 2008.

[30] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, 2010.

[31] Kavita Ganesan, ChengXiang Zhai, and Evelyne Viegas. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 869–878, 2012.

[32] Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, 2010.

[33] Andrea Glaser and Hinrich Schütze. Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 276–285, Avignon, France, April 2012.

[34] A. Go, R. Bhayani, , and L. Huang. Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University, 2009.

[35] S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 20–29, 2011.

[36] Narendra Gupta, Giuseppe Di Fabbrizio, and Patrick Haffner. Capturing the stars: Predicting ratings for service and product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 36–43, Los Angeles, California, June 2010.

[37] B. Han and T. Baldwin. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 368–378, 2011.

[38] Sanda Harabagiu and Andrew Hickl. Relevance modeling for microblog summarization. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[39] Sanda Harabagiu and Andrew Hickl. Relevance Modeling for Microblog Summarization. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[40] Ahmed Hassan, Dragomir R. Radev, Junghoo Cho, and Amruta Joshi. Content based recommendation and summarization in the blogosphere. In *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM)*, 2009.

[41] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA, 2004. ACM.

[42] Meishan Hu, Aixin Sun, and Ee-Peng Lim. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the Sixteenth ACM conference on Conference on Information and Knowledge Management (CIKM)*, pages 901–904, 2007.

[43] A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, and S. Gindl. Visualizing contextual and dynamic features of micropost streams. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[44] David Inouye and Jugal K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *SocialCom/PASSAT*, pages 298–306, 2011.

[45] Elham Khabiri, James Caverlee, and Chiao-Fang Hsu. Summarizing user-contributed comments. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[46] Hyun Duk Kim and ChengXiang Zhai. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 385–394, 2009.

[47] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 100–107, 2006.

[48] Patrick Lai. Extracting Strong Sentiment Trends from Twitter. *http://nlp.stanford.edu/courses/cs224n/2011/reports/patlai.pdf*, 2010.

[49] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–11628, August 2005.

[50] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, 2004.

[51] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web (WWW '05)*, pages 342–351, New York, NY, USA, 2005. ACM.

[52] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, pages 131–140, 2009.

[53] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1:309–317, 1957.

[54] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April 1958.

[55] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 2012.

[56] Inderjeet Mani. *Automatic Summarization*. John Benjamins Publishing Company, 2001.

[57] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 Conference on Human Factors in Computing Systems (CHI)*, pages 227–236, 2011.

[58] Richard McCreadie, Craig Macdonald, and Iadh Ounis. Identifying top news using crowdsourcing. *Information Retrieval*, pages 1–31, 2012. 10.1007/s10791-012-9186-z.

[59] B. Meyer, K. Bryan, Y. Santos, and B. Kim. Twitterreporter: Breaking news detection and visualization through the geo-tagged twitter network. In *Proceedings of the ISCA 26th International Conference on Computers and Their Applications*, pages 84–89, 2011.

[60] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, 2004.

[61] Shamima Mithun and Leila Kosseim. Summarizing blog entries versus news texts. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 1–8, Borovets, Bulgaria, September 2009.

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

[63] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *Web Information Systems Engineering*, pages 539–553, 2009.

[64] Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 670–679, 2011.

[65] Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.

[66] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, pages 145–152, 2004.

[67] A. Pak and P. Paroubek. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 436–439, 2010.

[68] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1), 2008.

[69] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on EMNLP*, pages 79–86, 2002.

[70] Bo Pang and Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, 2005.

[71] Michael J. Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, 2010.

[72] M. Pennacchiotti and A.M. Popescu. A Machine Learning Approach to Twitter User Classification. In *Proceedings of ICWSM 2011*, pages 281–288, 2011.

[73] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the 2009 ACM Conference on Recommender Systems*, pages 385–388, 2009.

[74] Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer*, 42(3):33–40, 2009.

[75] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 339–346, Vancouver, Canada, 2005.

[76] Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. MEAD — A platform for multidocument multilingual text summarization. In *Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004.

[77] Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, November 2004.

[78] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[79] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011.

[80] H. Saggion, K. Bontcheva, and H. Cunningham. Robust Generic and Query-based Summarisation. In *Proceedings of the European Chapter of Computational Linguistics (EACL), Research Notes and Demos*, 2003.

[81] A. Scharl and A. Weichselbraun. An automated approach to investigating the online media coverage of US presidential elections. *Journal of Information Technology and Politics*, 5(1):121–132, 2008.

[82] Frank Schilder, Ravikumar Kondadadi, Jochen L. Leidner, , and Jack G. Conrad. Thomson reuters at tac 2008: Aggressive filtering with fastsum for update and opinion summarization. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, pages 396–405, 2008.

[83] D.A. Shamma, L. Kennedy, and E.F. Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? In *Proceedings of CSCW 2010*, 2010.

[84] B. Sharifi, M. A. Hutton, and J. Kalita. 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, Los Angeles, California, June 2010.

[85] X. Song, Y. Chi, K. Hino, and B. L. Tseng. Summarization system by identifying influential blogs. In *Proceedings of ICWSM*, pages 325–326, 2007.

[86] Josef Steinberger, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, and Vanni Zavarella. Jrcs participation at tac 2011: Guided and multilingual summarization tasks. In *Proceedings of the Text Analysis Conference (TAC) 2011*, 2011.

[87] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(September 2010):1–41, 2011.

[88] Ivan Titov and Ryan Mcdonald. A joint model of text and aspect ratings for sentiment summarization. In *Proc. ACL-08: HLT*, pages 308–316, 2008.

[89] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, pages 417–424, Morristown, NJ, USA, July 2002.

[90] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang, and S. D. Lin. IMASS: An Intelligent Microblog Analysis and Summarization System. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 133–138, Portland, Oregon, 2011.

[91] W Wu, B Zhang, and M Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 689–692, 2010.

[92] W. Xin, Z. Jing, J. Jing, H. Yang, S. Palakorn, W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E. P. Lim, and X. Li. Topical keyphrase extraction from Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 379–388, 2011.

[93] Rui Yan, Mirella Lapata, and Xiaoming Li. Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 516–525, Jeju Island, Korea, 2012.

[94] Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49, March 2007.

[95] Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, 2011.

[96] F. Zanzotto, M. Pennaccchiotti, and K. Tsioutsiouliklis. Linguistic Redundancy in Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 659–669, Edinburgh, UK, 2011. Association for Computational Linguistics.

[97] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR)*, pages 338–349, 2011.

[98] Liang Zhou and Eduard Hovy. On the summarization of dynamically introduced information: Online discussions and blogs. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 237–242, 2006.

[99] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. Towards real-time summarization of scheduled events from twitter streams. *CoRR*, abs/1204.3731, 2012.