What's the Issue Here?: Task-based Evaluation of Reader Comment Summarization Systems

Emma Barker, Monica Paramita, Adam Funk, Emina Kurtic,
Ahmet Aker, Jonathan Foster, Mark Hepple and Robert Gaizauskas
Department of Computer Science
University of Sheffield, UK
{initial.surname}@sheffield.ac.uk

Abstract

Automatic summarization of reader comments in on-line news is an extremely challenging task and a capability for which there is a clear need. Work to date has focussed on producing extractive summaries using well-known techniques imported from other areas of language processing. But are extractive summaries of comments what users really want? Do they support users in performing the sorts of tasks they are likely to want to perform with reader comments? In this paper we address these questions by doing three things. First, we offer a specification of one possible summary type for reader comment, based on an analysis of reader comment in terms of issues and viewpoints. Second, we define a task-based evaluation framework for reader comment summarization that allows summarization systems to be assessed in terms of how well they support users in a time-limited task of identifying issues and characterising opinion on issues in comments. Third, we describe a pilot evaluation in which we used the task-based evaluation framework to evaluate a prototype reader comment clustering and summarization system, demonstrating the viability of the evaluation framework and illustrating the sorts of insight such an evaluation affords.

Keywords: Reader comment summarization, task-based evaluation, social media argumentation

1. Introduction

A common feature of on-line news sites is *reader comment* – a facility whereby readers of a news story can engage in conversation with each other, discussing aspects of or reactions to a news story. Often, news stories attract hundreds or even thousands of reader comments within a relatively short period of time. The problem then arises of *how to make sense* of this sprawling, multi-threaded conversation.

Intuitively, what one wants is some kind of summary or overview of the conversation, with the option of "drilling down" for more details. Since generating such an overview manually for every news story is clearly infeasible, automatic summarization offers promise here. And indeed, several authors have already proposed systems for summarizing reader comments (e.g. (Khabiri et al., 2011; Ma et al., 2012; Llewellyn et al., 2014)). Their systems are broadly similar, first topically clustering comments, then ranking comments within clusters and finally selecting top-ranked comments to form an extractive summary.

This extraction-oriented work begins with the assumption that topically grouped and ranked comments or extractive summaries are to likely to be useful to an end user. But it is true? To our knowledge there has been no attempt to investigate what an end user might want in a summary of reader comments. Furthermore, the evaluations proposed so far, despite in several cases being called user studies, are not task-based evaluations that might let us understand how well systems are meeting user needs.

A different, but promising, line of work, not yet deployed in summarization systems, is that on argument mining. Much of reader comment is argumentative and one appealing type of summary is one that would summarise the main points of contention in comments, something it is not clear an extractive summary could do. Work by e.g. Ghosh et al., (2014) Habernal et al. (2014) and Swanson et al. (2015) amongst others, focuses on defining and identifying key argumentative units and their relations. They mention summarization of argumentative texts as one potential application of their work. However, they do not specify what an end-user summary of reader comment on news might be like.

In this paper we make three contributions to advancing work in this area. First, we offer a specification of one possible summary type for reader comment based on the notions of *viewpoint* and *issue*, which we define below (Section 2.). Second, we propose a task-based evaluation framework in which users are offered access to the comments via alternative interfaces, some of which include summarization system outputs, and are asked to complete a short, time-limited writing task that requires understanding the comments (Section 3.). Third, we describe a pilot evaluation in which we used the task-based evaluation framework to evaluate a prototype reader comment clustering and summarization system, demonstrating the viability of the evaluation framework and illustrating the sorts of insight such an evaluation affords (Section 4.). In Section 5. we discuss related work and in Section 6. we conclude.

2. The Overview Summarization Task

In this section we specify one possible summarization task for a set of reader comments on a news article. We refer to this task as the *conversation overview* task. Before describing this summary type, we introduce some terminology and an informal framework for describing some aspects of the discourse structure of comment (our thinking here has been influenced in part by work in the argument mining community discussed further in Section 5.2.).

2.1. Discourse Structure of Reader Comment

Reader comments are typically presented in online news in association with a particular news article. Comments

Rubbish? Bury council votes to collect wheelie bins just once every three weeks
Locals fear the new move will lead to an increase in fly-tipping and attract foxes and vermin, but the council insists it will make the borough more environmentally friendly. Is it just a desperate cost cutting measure? . . .
A council in Greater Manchester is to be the first in England to start collecting wheelie bins only once every three weeks, scrapping the current fortnightly collection. The controversial decision was unanimously passed by councillors in Bury on Wednesday night, despite fears fly tipping would increase. One councillor who voted for the motion accused her opponents of "scaremongering" after they warned rubbish would pile up and attract vermin. Another argued the money saved could be spent on more social workers.
It affects the grey bins used for general household waste which can't be recycled . . . The Labour-run council claims the move is part of a strategy to turn Bury into a "zero waste borough", boost recycling and save money on landfill fees . . . many residents feel it is simply a desperate cost saving measure, after the town hall was told to make more than £32m of cuts over the next two years . . .

Id	Poster	Reply	Comment
1	A		I can't see how it won't attract rats and other vermin. I know some difficult decisions have to be made with cuts to funding, but this seems like a very poorly thought out idea.
2	B	2 → 1	Plenty of people use compost bins and have no trouble with rats or foxes.
3	C	3 → 2	If they are well-designed and well-managed- which is very easily accomplished. If 75% of this borough composted their waste at home then they could have their bins collected every six-weeks. It's amazing what doesn't need to be put into landfill.
4	D	4 → 2	Composting is for waste food. The black bin is for stuff that usually cannot be composted such as packaging - that comment doesn't solve or really relate to this. If the council randomly decides what services it will offer can we randomly decide how much we want to pay?
5	B	5 → 4	*The black bin is for stuff that usually cannot be composted such as packaging* Foxes, rats and other vermin don't really go looking for packaging, do they?
6	E	6 →1	It won't attract vermin if the rubbish is all in the bins. Is Bury going to provide larger bins for families or provide bins for kitchen and garden waste to cut down the amount that goes to landfill?
7	F	7 → 1	Expect Bury to be knee deep in rubbish by Christmas it's a lame brained Labour idea and before long it'll be once a month collections. I'm not sure what the rubbish collectors will be doing if there are any. We are moving back to the Middle Ages, expect plague and pestilence.
8	D	8 → 5	*Foxes, rats and other vermin don't really go looking for packaging, do they?* No but they may be drawn to the smells from used nappies, or leftover food on waste packaging for example . . .
9	E	9 → 5	They do if the packaging once held things like bacon, for example. There might not be any food, but they'll still make a huge mess trying to find it.

Figure 1: Part of a news article (top) and comments responding to it (bottom)[1]

are posted in a temporally ordered sequence of *threads*. Threads are temporally ordered sequences of *comments* in which each comment except the first is a reply to exactly one earlier comment in the thread, but where there may be many replies to each comment.

Within the constraints of this structure, comments are a form of *multi-party conversation* in which participants exchange views and opinion. Threads may be freely initiated, and they are not always topically coherent. Frequently, the same topic may be addressed by many threads, and a single thread may address many topics.

We treat the original article plus all comments as a single local discourse context and the task is to summarise the conversation within this discourse space. Of course, this space connects with a broader conversational discourse that is taking place within an entire population, perhaps across a big topic like "Britain's continued membership in the EU". But the task we focus on here is providing a summary of what emerges in the local discussion only.

In a nutshell our view is that reader comments address *issues*, readers hold *viewpoints* on issues and that comments make *assertions*, which serve many purposes including directly expressing a viewpoint. Of course comments may also have a social or pragmatic function, e.g. jokes, but we are primarily interested in semantic content, i.e. the assertions that are made and their role in expressing viewpoints on issues. We expand on these terms as follows:

Assertions A comment may comprise one or more assertions, i.e. propositions that the commenter believes to be true. Each assertion has a particular role in the local discourse, i.e. one assertion has a function in relation to other assertions. We find relations both between assertions made within a comment and between assertions made in different comments and assertions in the article. There is typically a central or primary assertion within a comment. Some key relations include the following: *grounds* (provides evidence/grounds to support another); *background* (provides background to another); *consequences* (indicates the possible consequences of another); *corrects* (corrects a detail of a prior assertion); *rebuttal* (a reader expresses disagreement with or rejects the validity of another claim); *agree/confirm* (a reader indicates agreement or confirms validity of another claim); *questions* (a reader questions another claim – why is that? where is that? who is that? is that true?)

Viewpoints or stance Disagreement or contention between comments is a pervasive feature of reader comment and news. When an assertion made by one comment is contradicted by i) an assertion expressed in another comment, or ii) an assertion reported in or entailed by something reported in the news article, each opposed assertion expresses a *viewpoint* or *stance*. In other words a viewpoint or stance

[1] Source: `www.theguardian.com/uk-news/the-northerner/2014/jul/17/rubbish-bury-council-votes-to-collect-wheelie-bins-just-once-every-three-weeks`

can be determined when an assertion contends with another assertion in the comment stream or when a commenter is "arguing with" an assertion reported in the original article. It follows that whether or not an assertion expresses a viewpoint is an emergent property of the discourse and only relative to the discourse; it is not an inherent feature of the assertion itself.

Issues Implicitly related to notion of viewpoint is that of *issue*. We can think of an issue as a question or problem, to which there are multiple contending answers. The space of possible answers is the set of related but opposed viewpoints expressed in the comment set. I.e. an issue is that which a viewpoint or stance is a viewpoint *on*.

Issues may typically be expressed via a "whether or not" type expression, e.g. the issue of *whether or not to lower the drinking age*, a question with binary opposites, *Should Britain leave the EU? (yes, no)*; via a "which is the best X?"-type expression when there are more than two opposed alternatives, e.g. the issue of *which was the best film of 2015?*. However, issues are rarely explicitly articulated in reader comments, (or the news article, see e.g. Figure 1). Rather, as the dialogue evolves, a set of assertions made by commenters may indicate a space of alternative, opposed viewpoints, and an issue can then be recognised/articulated. Sub-issues may emerge within the discussion around an issue, i.e. there is a recursive nature to issues. E.g. when evidence proposed as support for a viewpoint on an issue is contended by another comment, the two 'new' contending viewpoints may be addressed by further comments, and thus a new issue can be recognised. Finally, it is worth noting that threads and issues are not the same thing: comments addressing any one issue may occur in multiple threads and any one thread may contain comments relating to multiple issues.

Example To illustrate the concepts of assertions, viewpoints and issues in news and reader comment we can refer to the example text in Figure 1. This shows an extract from a news article about the controversy surrounding a council's decision to reduce bin collection and examples from a thread of comments posted in response to the news report. The article text reports opposing viewpoints, some people (e.g. 'councillors') supported this decision while others (e.g. 'locals', 'residents', 'opponents') objected. We can summarise the different arguments put forward in support of the respective positions and identify the main issue in the news as follows:

Issue: is the decision to reduce bin collection a good one?

Viewpoint: reducing bin collection is a good decision

Grounds: will be environmentally friendly and encourage recycling/composting, saved money could go to other services e.g. social workers.

Viewpoint: reducing bin collection is a bad decision

Grounds: will attract vermin, increase fly tipping, is a cost cutting measure with no benefits.

The comments proceed by expressing different positions on this issue: comments 1, 4, 7 show direct opposition to the decision, e.g. "it's a poorly thought out idea"; "it's a lame brained Labour idea". By contrast, comment 3 supports the position that less frequent collection is a good idea, "if we compost more we have less rubbish and we wont need bin collection regularly."

Comment 1 makes the assertion, also reported in the article, that less frequent collection will attract vermin "can't see how it won't attract rats and other vermin". Comments 2 and 6 contend this view, 2 citing an example of compost bin users "having no trouble with rats or foxes", and 6, "it won't attract vermin if its all in the bins". Note also the report in the article that a councillor accused opponents of "scaremongering" after they warned about vermin. So, a sub-issue can be clearly recognised here, which may be articulated as "whether or not less frequent bin collection will attract vermin".

To complicate things further, comments 4,8,9 take issue with the assertion implicit in 2, i.e. that compost bins which contain food don't attract vermin so why should grey bins? 4,8,9 together maintain that the rubbish in the grey bins such as food packaging might have remains of e.g. bacon, pizzas, etc. and this won't be composted, so grey bins are a different case and can attract vermin.

This last example doesn't attract many comments and so is better seen as a minor issue, e.g. "does the rubbish found in grey bins attract vermin?", as opposed to a larger issue. Note that while comments 4,8 and 9 together express support for the view that packaging in grey bins can attract vermin, they do not necessarily support the view that less frequent bins will attract vermin.

Finally, as this example clearly shows, the structure of viewpoints and issues in reader comment is very complex. The problem of complexity is further compounded by the volume of comment that typically follows an article. It follows that any summarization task that exploits such structure will be hard to carry out.

2.2. The Summarization Task

Given this informal account of assertion, viewpoint and issue in comment and news, we now offer our specification of what a conversational overview summary of reader comments should contain. **Ideally, a summary should:**

1. **Identify and articulate main issues in the comments.** Main issues are those receiving proportionally the most comments. They should be prioritized for inclusion in a space-limited summary.
2. **Characterise opinion on the main issues.** To characterise opinion on an issue typically involves:
   - identifying alternative viewpoints;
   - indicating the grounds given to support a viewpoint;
   - aggregation: indicating how opinion was distributed across different issues, viewpoints and grounds, using quantifiers or qualitative expressions e.g "the majority discussed x";
   - indicating where there was consensus or agreement among the comment;
   - indicating where there was disagreement among the comment;

We presented the proposed summary specification to a range of reader comment users, including comment readers, posters, journalists and news editors and received very positive feedback via a questionnaire.[2]
Example (6), which summarises comment on a news story about a heat-wave in the UK, illustrates the characterisation of one issue.

(6) *The majority of comments discussed what was the best way to stay cool in the heat. Some said that air-conditioning was useful and the best way to stay cool in high temperatures; but others said air-con units are unneccessary in the UK and preferred to use fans. A few objected to air-con saying it is bad for the environment, too expensive to run and very noisy. Whereas fans were said by some to be cheap.*

3. A Task-based Evaluation Framework

Systems developed to produce conversation overview summaries will need to be evaluated. Intrinsic evaluation of such summaries, either by direct assessment or by comparison against a set of reference summaries, is one potential means of assessment, the pros and cons of which are well documented. But, given the difficulty of the conversation overview task and the likelihood that for the foreseeable future outputs will fall short of any reference summary standard, an alternative approach to evaluation is to ask: can automatic summaries and clusters help a user in the context of intended use? To address this we require: a practical evaluation task for users to carry out; a software platform with articles, comments and system outputs presented in a navigable interface, such that a user may interact with the system outputs in the context of the full comment stream; and a set of metrics to allow us to assess how different outputs in the context of such an interface might help users complete the task.

3.1. Evaluation Tasks

We propose the following series of tasks for users to carry out in such an evaluation:

1. Simplified overview task: first, we ask participants to imagine they are a user wanting to make sense of a comment conversation in a short period of time, e.g. a coffee break; we then provide users with a system and a topic (an article and comment set); allow a set time for reading over news and comment (e.g. 2 mins) and then ask users to: (1) identify four main issues in the discussion and (2) characterise opinion on a given issue in a set time (e.g. 10 mins) in accordance with our definitions. Participants are asked to carry out the constrained overview task for multiple system-topic combinations (see below).
2. Post task questionnaire: we ask participants to rate and compare the usefulness of the system(s) and system components in the context of completing the tasks, on a 5-point scale and include an option for written feedback.
3. Finally, in a guided group discussion we invite participants to comment on their experience during the tasks and on using the different systems/components.

This three-part protocol provides three complementary sets of results. To compare systems, we can now design experiments with any number of different system-variants, involving participants and topics as required, to control for topic effects and individual user differences. We then use the results of the protocol with each task instance to compare how, and to what extent, the different systems help users in carrying out the constrained overview task.

3.2. Metrics

Giving assessors the source comments and the news article, we assess written responses obtained from the overview questions using a novel, graded scheme. Each issue is scored on a 4 point scale that takes account of criteria such as evidence/accuracy and clarity of expression (how clearly is the issue articulated?). Characterisation of opinion is scored on a graded 6 point scale, based on criteria of coverage, representing quantities and accuracy.

Q1: "Identify Four Issues"
To assess the quality of participant responses to this task we used a four point scale, ranging from 0-3. Judges assigned an individual score to each of 4 issues. The 4 point scale takes account of criteria including "*evidencing*" (i.e., is there evidence for the issue in the comments? is it an accurate description of a "main issue" in the comments?); and "*clarity of expression*" (how clearly is the issue articulated?). Guidelines for assessing Q1 are shown in Table 1.

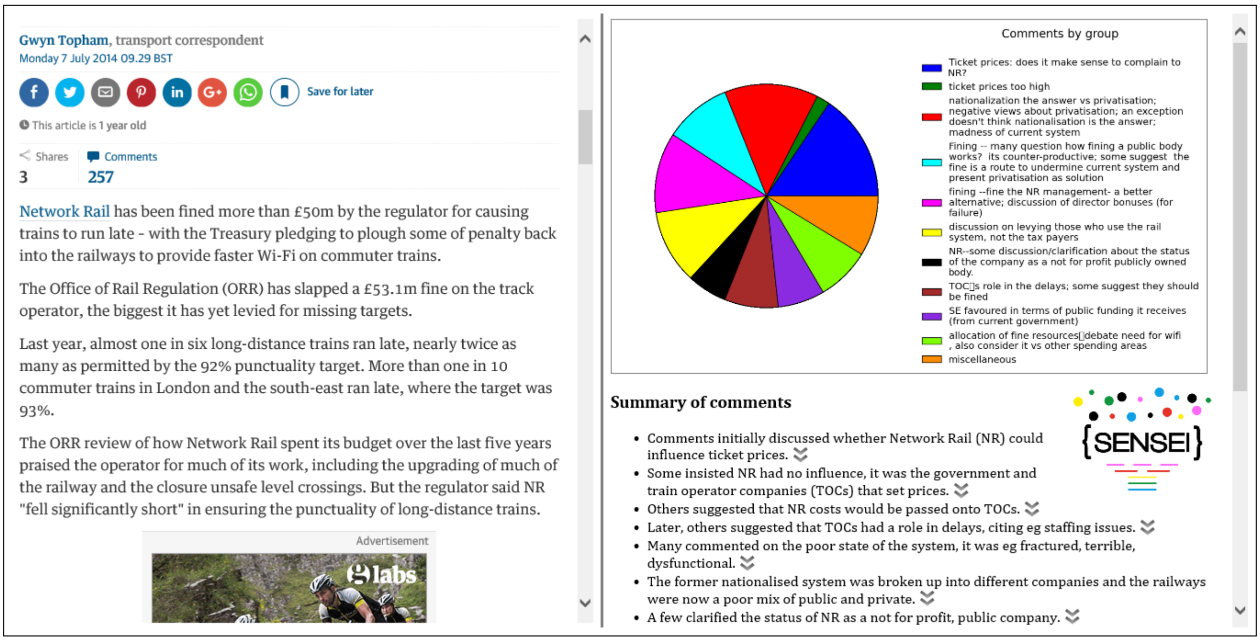
Table 1: Guideline for Q1 scores

Score 0: No issue given or issue given but no evidencing apparent; a well-articulated issue with no evidencing in the comment would receive a score of 0.
Score 1: The issue is expressed poorly, but some content is indicated and the comments can be seen to address it, for example, for a response "ticket prices" there is evidence of people talking about different things to do with ticket prices in the comments. Or, the issue is more clearly articulated e.g. as a proposition or as a question, but is poorly evidenced, e.g. only 1 or 2 comments discuss the issue.
Score 2: The issue is adequately expressed e.g. "fining directors", but one could imagine the space of possible positions being more clearly indicated e.g.'would fining directors be an effective way of ensuring trains run on time". The issue should be of sufficient clarity to assess evidence or strength of support in the comments, which should be good or satisfactory. Or, a well-articulated issue but with a low level of evidencing, say 2-3 comments, or when there were many other candidate issues to choose from, which were much more significantly discussed.
Score 3: The issue is clearly articulated/expressed; so it is straightforward to assess evidencing/strength of support, which is good (relative to the overall discussion in the comments).

Q2: "Characterizing Opinion"
To assess the quality of participant responses to this task we used a six point scale, ranging from 0-5, with maximum and minimum values given: 0 = no characterisation of opinion present, 5 = excellent characterisation of opinion in the

[2] Further details on the summary specification and the end-user survey on it can be found in SENSEI deliverable D1.2 "Report on Use Case Design and User Requirements" at: `http://www.sensei-conversation.eu/deliverables/`

Figure 2: The system interface



Gwyn Topham, transport correspondent
Monday 7 July 2014 09.29 BST

Save for later

This article is 1 year old

Shares	Comments
3	257

Network Rail has been fined more than £50m by the regulator for causing trains to run late - with the Treasury pledging to plough some of penalty back into the railways to provide faster Wi-Fi on commuter trains.

The Office of Rail Regulation (ORR) has slapped a £53.1m fine on the track operator, the biggest it has yet levied for missing targets.

Last year, almost one in six long-distance trains ran late, nearly twice as many as permitted by the 92% punctuality target. More than one in 10 commuter trains in London and the south-east ran late, where the target was 93%.

The ORR review of how Network Rail spent its budget over the last five years praised the operator for much of its work, including the upgrading of much of the railway and the closure unsafe level crossings. But the regulator said NR "fell significantly short" in ensuring the punctuality of long-distance trains.

Advertisement

Comments by group

- Ticket prices: does it make sense to complain to NR?
- ticket prices too high
- nationalization the answer vs privatisation; negative views about privatisation; an exception doesn't think nationalisation is the answer; madness of current system
- Fining -- many question how fining a public body works? its counter-productive; some suggest the fine is a route to undermine current system and present privatisation as solution
- fining --fine the NR management- a better alternative; discussion of director bonuses (for failure)
- discussion on levying those who use the rail system, not the tax payers
- NR--some discussion/clarification about the status of the company as a not for profit publicly owned body.
- TOC□s role in the delays; some suggest they should be fined
- SE favoured in terms of public funding it receives (from current government)
- allocation of fine resources□debate need for wifi , also consider it vs other spending areas
- miscellaneous

Summary of comments

- Comments initially discussed whether Network Rail (NR) could influence ticket prices.
- Some insisted NR had no influence, it was the government and train operator companies (TOCs) that set prices.
- Others suggested that NR costs would be passed onto TOCs.
- Later, others suggested that TOCs had a role in delays, citing eg staffing issues.
- Many commented on the poor state of the system, it was eg fractured, terrible, dysfunctional.
- The former nationalised system was broken up into different companies and the railways were now a poor mix of public and private.
- A few clarified the status of NR as a not for profit, public company.

response. An “excellent” rating requires an answer to include: good coverage of opinion on the issue, i.e. details of the different perspectives on the issue/the different sides to the argument; where there was consensus or not; some detail of the respective quantities of opinion; and the characterisation should be accurate, i.e. there should be evidence for the information given in the comments.

Post-task questionnaire and discussion

We analyze the free text and spoken responses gathered in the post task questionnaire and discussion using simple qualitative techniques. Data from the user ratings of the different systems/system components are summarised using simple statistics.

3.3. Interface

To carry out comparative evaluations of different systems we have developed a configurable interface with the following characteristics:

- It includes a baseline comment-only system, which presents threaded conversations in the way they typically appear in on-line news today.
- It takes as input comment clusters, labels for these clusters and summaries (either extractive or abstractive) and may contain links from summary sentences to the comment clusters from which they derive.
- It offers a text-based summary presentation mode in which the summary and a textual representative of each cluster (e.g. a cluster label or representative phrase or sentence) are displayed. If links to clusters are provided, summary sentences can be clicked to display the corresponding clusters. The textual representative of clusters can also be clicked, to display cluster comments.
- It offers a graphical summary presentation mode in which a piechart displays the clusters as wedges, whose size reflects the proportion of comments in the associated cluster, and where wedges are labelled with the textual representative of clusters. The wedges can be clicked to display cluster comments .
- Where a cluster has been expanded to display its comments, any comment may be clicked to show it in its original context in the comment stream.

This configurable interface allows an evaluation to be run where two or more systems are compared, the systems being differentiated either in terms of the algorithms they employ to carry out comment clustering, labelling and summarization or in terms of how they present the results of this underlying linguistic processing to users.

4. A Pilot Evaluation

4.1. Evaluation setup

We tested the full task protocol and interface in a pilot evaluation. Four participants, all post-graduates with experience in language technologies and using reader comment, each carried out two iterations of the task, each time using a different system/interface configuration:

S1 A baseline, presented just the reader comment facility used by *The Guardian* in current practice.

S2 Included both the current practice facility and sense-making components, consisting of a labelled pie chart indicating the relative size of comment clusters and a summary whose sentences were linked to underlying comment clusters. This interface is shown in Figure 2.

The clustering, cluster labelling and summarization outputs were produced by an early versions of a baseline extractive reader comment summarization system [3].

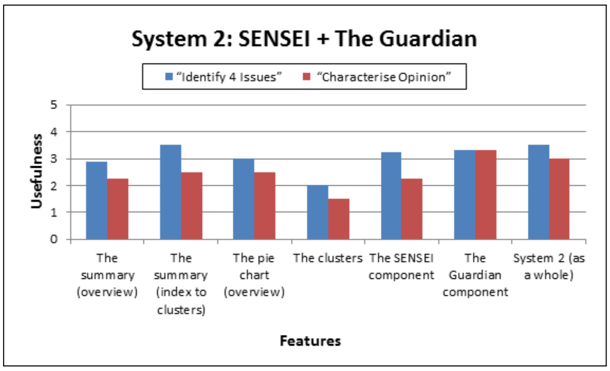
[3] Details of this system may be found in SENSEI deliverable D5.2 “Specification of Conversation Analysis Summarization Outputs” at: `http://www.sensei-conversation.eu/deliverables/`

Table 2: Evaluation results

System	Participant ID	Q1	Q2	Total Score
Article: “Network Rail” (T1)				
S1	P1	10	0	10 (58.8%)
S1	P4	11	3	14 (82.4%)
S2	P2	4	1	5 (29.4%)
S2	P3	5	2	7 (41.2%)
	Grand Total	30	6	36 (52.9%)
Article: “Heatwave” (T2)				
System	Participant ID	Q1	Q2	Total Score
S1	P2	10	0	10 (58.8%)
S1	P3	12	2	14 (82.4%)
S2	P1	9	4	13 (76.5%)
S2	P4	12	4	16 (94.1%)
	Grand Total	43	10	53 (77.9%)

Total score is also represented as a percentage of the maximum possible score. The max score per task is 17: 12 points for Q1 (4 questions × 3 points), plus 5 points for Q2. For each article, the max score is 68, i.e. 17 points for each of 4 participants

Figure 3: Post-task questionnaire results



There were two different topics, each comprising a news article and an associated set of 100 comments. Each participant used each system and each topic exactly once. We provided a short training session including a system demo and guidelines on the overview scenario and tasks.

4.2. Results

4.2.1. Overview questions

The evaluation results, shown in Table 2, allow us to gain insights on various research questions, as follows.

Was one topic easier to answer than another? The score for topic T2 (“Heatwave”) is higher than for T1 (“Network Rail”), at 77.9% vs 52.9%, reflecting the fact that T1 is a more complex topic with longer comments. Indeed, both questions received better scores for T2 than T1.

Was one content question easier than another? Overall, scores for Q1 were higher than those for Q2, suggesting that identifying issues is easier than characterising opinion.

Did different systems help with the different content questions? For Q1, System S1 scores much higher than S2 (89.6% vs 62.5%). For Q2, S2 scores higher than S1 (55% vs 25%). The performance of S1 on Q1 suggests that Guardian threads work quite well for identifying issues in comment. But the task might prove more difficult if the number of comments to summarize was increased.

Did one system help overall more than another? Comparing scores within system conditions, we find system scores to be broadly similar, i.e. 70.6% for S1, vs 60.3% for S2.

Did a particular system help with a particular topic? Scores for S1 were the same across topics, while scores for S2 were much higher for the simpler topic T2 (“Heatwave”) than the more complex T1.

Were some participants better at answering the questions than others? Results show notable differences between participants. Scores for P4 were consistently high, at 88.2% overall, while P2’s were low at 44.1%. Scores for P3 and P1 were fairly similar, at 61.8% and 67.6% respectively. The results suggest a need, in future work, for using more participants and topics, to counter effects of bias from a specific assignment of individuals to systems and topics.

4.2.2. Post-task questionnaire

In the post-task questionnaire, participants rated how useful different systems/system components were to completing the tasks, on a scale of 1 (not useful) to 5 (extremely useful). Participants found System 1, as a whole, to be useful in completing the tasks, especially the comment threads. Usefulness scores for System 2 (averaged over participants) are shown in Figure 3.

4.2.3. Guided group discussion

The discussion was broadly divided into 3 high level questions about participant experience of 1) the experimental tasks; 2) the systems and 3) an open question. The responses to the first two questions covered four aspects of the evaluation.

- the evaluation tasks: reading the news article, reading the comment sets using either of the interfaces, content questions (identify 4 issues and characterise opinions);
- the two systems for making sense of comments;
- their strategies in solving the tasks: which features of the systems were used for each of the two question types;
- general usefulness of the SENSEI system.

4.2.4. Summary

The three complementary sets of results allowed us to assess the protocol and to compare how, and to what extent, the different systems/components helped users to complete the two tasks. Whilst feedback on prototype suggests a need for further development, if outputs are to be helpful, the general interface design and direction of the technology, as guided by the overview task, was approved. Results also indicate that the protocol provides sufficient data to answer our research questions [4].

[4] Further details of the evaluation framework and piolot evaluation may be found in SENSEI deliverable D1.3 “Intermediate Evaluation” at: `http://www.sensei-conversation.eu/deliverables/`

5. Related Work

5.1. Summarising Reader Comment

As noted above in Section 1., a small number of authors have directly addressed the task of summarizing on-line conversations commenting on videos or news articles. To date, all have adopted broadly similar approaches, first topically clustering comments, then ranking comments within clusters and finally selecting top-ranked comments to form an extractive summary. We comment here on the task addressed and the evaluation methods used, skipping over details of the language processing techniques employed.

Khabiri et al. (2011) address the task of summarising comments relating to Youtube videos. They carry out just the first two stages of the general three stage process outlined above, i.e. topical clustering and ranking. To evaluate the comment ranking they first produced a gold standard resource by asking multiple annotators to identify which comments in the first 50 comments on a video were interesting and informative and assigning an informativeness score to each comment based on how many annotators judged the comment to be informative. System outputs were then scored against the human ranking derived from this comment scoring using the standard information retrieval measure of normalised discounted cumulative gain (NDCG). They used their gold standard resource and NDCG to evaluate alternative ranking approaches.

Ma et al. (2012) address the task of summarising reader comments in *Yahoo! News*, with a view to generating "an easy overview of all topics discussed in the comments". They carry out the full three stage process outlined above. A final summary of 15 comments is formed from the five top-ranked comments chosen from the three largest clusters. The results are evaluated through a user study assessing topic cohesion, topic diversity, and news relatedness, each on a 5-point scale. Three subjects used these criteria to assess six summarisation systems, reading summaries of 50 news articles for each system. Scores are aggregated across participants and news articles to produce overall scores, permitting conclusions to be drawn about which system performs best according to which criterion.

Llewellyn et al. (2014) address the task of summarising reader comments in *The Guardian* newspaper, again following the three stage process outlined above and finishing by selecting top ranked comments across multiple clusters to form a summary. To evaluate ranking methods they use a set of human-authored short summaries produced by *The Guardian* for selected comment sets as a gold standard. For each gold standard summary, subjects were asked to rank seven different system-generated summaries by comparison with it. System summaries were formed by taking the three top-ranked comments from each cluster, for seven different combinations of clustering and ranking techniques.

In our view, all of the above work is limited by assumptions made about the nature of the task and of the evaluations carried out. The authors only consider summaries comprising extracted reader comments or even just ranked lists of comments within topics. However, it is not clear that comments extracted from the comment stream, i.e. pulled out of their dialogic context, are either what summary readers might want or are even likely to make sense. While all three authors carry out evaluations that include users, these evaluations are focussed on giving insights into the relative merits of various technologies, e.g. they allow conclusions of the form "users prefer outputs produced by this ranking technique to that one". However, these evaluations do not provide any sense of whether the resulting summaries actually give users what they want or allow users to gain information that would enable to carry out some task. Our work addresses both these issues.

5.2. Argument Extraction

In recent years various authors have begun work on the problem of identifying units of argumentation in online discussion forums and reader comment. Ghosh et al. (2014), for example, propose an annotation scheme for identifying argument units in on-line interactions – which they call *call-outs* and *targets* – and relations between them. A call-out refers back to an earlier target and includes a stance and/or rationale with respect to the target. These notions are similar to our notion of viewpoint, sharing particularly the idea that they gain their character only as contention or interaction between participants takes place. Swanson et al. (2015) propose the task of identifying and extracting *argument facets* in text – phrases that express arguments about sub-issues in the context of an argument about a wider issue or topic (e.g. "gun control") and that can be understood without any additional context. They are interested in identifying facets and then grouping similar facets with a view to "producing argument summaries that reflect the range and type of arguments being made on a topic, over time, by citizens in public forums". Habernal et al. (2014) review a wide range of different argument annotation schemes and compare two on the task of annotating different text types, concluding that the choice of scheme depends on the nature of the application and the type of text to be annotated. They note the potential utility of the *Claim-Premise scheme* for automatic summarization by allowing arguments with similar claims or premises to be clustered together.

Although the precise details of argumentation models may vary between authors in terms of its elements, the terminology used to label them and the relations between them, there are a number of common assumptions: that identifying and extracting argument elements is an important task; that, once identified, similar elements can be collected together; and that such aggregation can be used to generate useful and informative summaries of the argumentative content. However, while progress has been made on low level annotation schemes for particular argumentative elements, no example summaries have been created and the precise form or character of an end user summary – whether in relation to a single issue across multiple documents or multiple issues that emerge within a single comment set (i.e. what we propose above) – is something the argument mining community has not yet addressed.

5.3. Task-based Evaluation

In this paper we have presented a task-based approach to summarisation system evaluation, also referred to as *extrinsic evaluation*, where a system is indirectly assessed by measuring the extent to which the system's outputs help a user to perform some task external to the system itself. The merits of such evaluations, which complement *intrinsic evaluation*, are well-recognised (Mani, 2001; Spärck Jones, 2007). Mani (2001) distinguishes two broad classes of tasks that have been explored for extrinsic evaluation of summarization systems: *relevance assessment* and *reading comprehension*. In the first summaries are assessed according to how well they support judgements about the relevance of document(s) they summarise to some task, while in the second summaries are assessed in terms of how well they are able to supply answers to a set of questions.

Clearly the evaluation we propose in this paper is a type of reading comprehension evaluation. The closest prior work to ours, also falling into this category, is McKeown et al.'s (2005) task-based evaluation of their *Newsblaster* multidocument summarisation system. In their study: "Four groups of subjects were asked to perform the same time-restricted fact-gathering tasks, reading news under different conditions: no summaries at all, single sentence summaries drawn from one of the articles, Newsblaster multi-document summaries, and human summaries." Once subjects had read whatever news source they were given, they were asked to write a brief report in response to three specific questions about the chosen news scenario. These reports were then manually assessed using the Pyramid method (Nenkova et al., 2007), where Pyramids were constructed for each question using the reports written by all study participants other than the one being assessed. The authors carried out an analysis of variance on the results to study the impact of the type of summary on report quality and also included the factors report writer, report topic and question in the model to estimate their contribution to the report quality. They also gave users a multi-question user satisfaction questionnaire on completion of the task.

Our approach has many similarities to this. In particular we too ask subjects to answer questions about source documents under different summarisation conditions. We also follow up the question answering task with a user satisfaction survey. The most notable differences in the two studies are their scale – McKeown et al. used 45 participants while we used four – and the method for assessing subjects' responses to the question answering task. Regarding scale, recall that our study was just a pilot; we plan a substantially larger scale exercise. Regarding assessment of subjects' responses, we used one human assessor, who had access to the original news articles and comments and to at least two human authored overview summaries (as specified in Section 2.2.), while they used Pyramid evaluation as described above. We believe our approach should be made more robust by using more than one human assessor, but are not convinced that the substantial resource commitment that a full Pyramid evaluation entails is necessary to achieve robust results and valuable insights into summary utility.

6. Conclusion and Future Work

In this paper we have made a proposal for what an overview summary of reader comment conversation should contain, based on an analysis of the nature of reader comment conversations. We then described a reusable framework for task-based evaluation of systems seeking to generate such overview summaries and finished by describing a pilot evaluation we carried out to assess the viability of our evaluation framework and to gain feedback on early versions of reader comment technologies we are developing. Our pilot has shown that our task-based evaluation methodology is indeed viable and that significant insights into both underlying language processing technologies and summary presentation techniques can be gained using it.

We intend to run the same evaluation protocol on more refined versions of our technology and with significantly larger numbers of topics and participants. We will also explore the possibility of interesting others in our approach and using the protocol as part of a shared task challenge on reader comment summarization.

7. Acknowledgements

This work was carried out as part of the EU-funded FP7 SENSEI project, grant number FP7-ICT-610916, and the authors would like to European Council for their support.

8. References

Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). Analyzing argumentative discourse units in online interactions. In *Proc. of the First Workshop on Argumentation Mining*, pages 39–48.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In *Proc. of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39.

Khabiri, E., Caverlee, J., and Hsu, C.-F. (2011). Summarizing user-contributed comments. In *Proc. of The Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, pages 534–537, Barcelona.

Llewellyn, C., Grover, C., and Oberlander, J. (2014). Summarizing newspaper comments. In *Proc. of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, June 1-4, 2014.*

Ma, Z., Sun, A., Yuan, Q., and Cong, G. (2012). Topic-driven reader comments summarization. In *Proc. of the 21st ACM international Conference on Information and Knowledge Management*, CIKM '12, pages 265–274.

Mani, I. (2001). Summarization evaluation: An overview. In *Proc. of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*, pages 77–85.

McKeown, K., Passonneau, R. J., Elson, D. K., Nenkova, A., and Hirschberg, J. (2005). Do summaries help? In *Proc. of the 28th SIGIR Conference*, pages 210–217.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The Pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4, May.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proc. of the SIGDIAL 2015 Conference*, pages 217–226.