The SENSEI Annotated Corpus: Human Summaries of Reader Comment Conversations in On-line News

Emma Barker, Monica Paramita, Ahmet Aker, Emina Kurtic,
Mark Hepple and **Robert Gaizauskas**
University of Sheffield, UK
e.barker@ m.paramita@ ahmet.aker@ e.kurtic@
m.r.hepple@ r.gaizauskas@ sheffield.ac.uk

Abstract

Researchers are beginning to explore how to generate summaries of extended argumentative conversations in social media, such as those found in reader comments in on-line news. To date, however, there has been little discussion of what these summaries should be like and a lack of human-authored exemplars, quite likely because writing summaries of this kind of interchange is so difficult. In this paper we propose one type of reader comment summary – the *conversation overview* summary – that aims to capture the key argumentative content of a reader comment conversation. We describe a method we have developed to support humans in authoring conversation overview summaries and present a publicly available corpus – the first of its kind – of news articles plus comment sets, each multiply annotated, according to our method, with conversation overview summaries.

1 Introduction

In the past fifteen years there has been a tremendous growth in on-line news and, associated with it, the new social media phenomenon of on-line reader comments. Virtually all major newspapers and news broadcasters now support a reader comment facility, which allows readers to participate in *multi-party conversations* in which they exchange views and opinion on issues in the news.

One problem with such conversations is that they can rapidly grow to hundreds or even thousands of comments. Few readers have the patience to wade through this much content. One potential solution is to develop methods to summarize comment automatically, allowing readers to gain an overview of the conversation.

In recent years researchers have begun to address the problem of summarising reader comment. Broadly speaking, two main approaches to the problem have been pursued. In the first approach, which might be described as *technology-driven*, researchers have proposed methods to automatically generate summaries of reader comment based on combining existing technologies (Khabiri et al., 2011; Ma et al., 2012; Llewellyn et al., 2014). These authors adopt broadly similar approaches: first reader comments are topically clustered, then comments within clusters are ranked and finally one or more top-ranked comments are selected from each cluster, yielding an extractive summary. A significant weakness of such summaries is that they fail to capture the essential argument-oriented nature of these multi-way conversations, since single comments taken from topically distinct clusters do not reflect the argumentative structure of the conversation.

In the second approach, which might be characterised as *argument-theory-driven*, researchers working on argument mining from social media have articulated various schemes defining argument elements and relations in argumentative discourse and in some cases begun work on computational methods to identify them in text (Ghosh et al., 2014; Habernal et al., 2014; Swanson et al., 2015; Misra et al., 2015). If such elements and relations can be automatically extracted then they could serve as the basis for generating a summary that better reflects the argumentative content of reader comment. Indeed, several of these authors have cited summarization as a motivating application for their work. To the best of our knowledge, however, none have proposed how, given an analysis in terms of their theory, one might produce a summary of a full reader comment set.

Proceedings of the SIGDIAL 2016 Conference, pages 42–52,
Los Angeles, USA, 13-15 September 2016.

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 →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? Many people won't fill the bins in 3 weeks - even when there was 5 of us here, we would have just about managed.
5	E	5 → 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.
6	F		Are they completely crazy? What do they want a new Plague?
7	G	7 →6	Interesting how you suggest that someone else is completely crazy, and then talk about a new plague.
8	H	8 →7	Do you think this is a good idea? We struggle with fortnightly collection. This is tantamount to a dereliction of duty. What are taxpayers paying for? I doubt anyone knew of this before casting their vote.
9	I	9→8	I think it is an excellent idea. We have fortnightly collection, and the bin is usually half full or less[family of 5].. Since 38 of the 51 council seats are held by Labour, it seems that people did vote for this. Does any party offer weekly collections?
10	G	10 →8	I don't think it's a good idea. But..it won't cause a plague epidemic.

Figure 1: Comments responding to a news article announcing reduced bin collection in Bury. Full article and comments at: `http://gu.com/p/4v2pb/sbl`.

In our view, what has been lacking so far is a discussion of and proposed answer to the fundamental question of what a summary of reader comments should be like and human-generated exemplars of such summaries for real sets of reader comments. A better idea of the target for summarisation and a resource exemplifying it would put the community in a better position to choose methods for summarisation of reader comment and to develop and evaluate their systems.

In this paper we make three principal contributions. First, after a brief discussion of the nature of reader comment we make a proposal about one type of informative reader comment summary that we believe would have wide utility. Second, we present a three stage method for manually creating reference summaries of the sort we propose. This method is significant since the absence to date of human-authored reader comment summaries is no doubt due to the very serious challenge of producing them, something our method alleviates to no small degree. Third, we report the construction and analysis of a corpus of human-authored reference summaries, built using our method – the first publicly available corpus of human-authored reader comment summaries.

2 Summaries of Reader Comments

What should a summary of reader comment contain? As Spärck-Jones (2007) has observed, what a summary should contain is primarily dependent on the nature of the content to be summarised and the use to which the summary is to be put. In this section we first make a number of observations about the character of reader comments and offer a specification for a general informative summary.

2.1 The Character of Reader Comments

Figure 1 shows a fragment of a typical comment stream, taken from reader comment responses to a *Guardian* article announcing the decision by Bury town council to reduce bin collection to once every three weeks. While not illustrating all aspects of reader comment interchanges, it serves as a good example of many of their core features.

Comment sets are typically organised into *threads*. Every comment is in exactly one thread and either initiates a new thread or replies to exactly one comment earlier in a thread. This gives the conversations the formal character of a set of trees, with each thread-initial comment being the root node of a separate tree and all other comments being either intermediate or leaf nodes, whose parent is the comment to which they reply. While threads may be topically cohesive, in practice they rarely are, with the same topic appearing in multiple threads and threads drifting from one topic onto another (see, e.g. comments 5 and 6 in Figure 1 both of which cite plague as a likely outcome of the new policy but are in different threads).

Our view, based on an analysis of scores of comment sets, is that reader comments are primarily argumentative in nature, with readers making *assertions* that either (1) express a *viewpoint* (or

stance) on an *issue* raised in the original article or by an earlier commenter, or (2) provide *evidence* or grounds for believing a viewpoint or assertion already expressed. Issues are questions on which multiple viewpoints are possible; e.g., the issue of whether reducing bin collection to once every three weeks is a good idea, or whether reducing bin collection will lead to an increase in vermin. Issues are very often implicit, i.e not directly expressed in the comments (e.g., the issue of whether reducing bin collection will lead to an increase in vermin is never explicitly mentioned yet this is clearly what comments 1-4 are addressing). A fuller account of this issue-based framework for analysing reader comment is given in Barker and Gaizauskas (2016).

Aside from argumentative content, reader comments exhibit other features as well. For example, commenters may seek clarification about facts (e.g. comment 4 where the commenter asks *Is Bury going to provide larger bins for families ...?*). But these clarifications are typically carried out in the broader context of making an argument, i.e. advancing evidence to support a viewpoint. Comments may also express jokes or emotion, though these too are often in the service of advancing some viewpoint (e.g. sarcasm or as in comments 4 and 6 emotive terms like *lame-brained* and *crazy* clearly indicating the commenters' stances, as well as their emotional attitude).

2.2 A Conversation Overview Summary

Given the fundamentally argumentative nature of reader comments as sketched above, one type of summary of wide potential use is a generic informative summary that aims to provide an overview of the argument in the comments. Ideally, such a summary should:

1. **Identify and articulate the 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 viewpoints; aggregating – 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 this proposed summary type to a range of reader comment users, including comment readers, posters, journalists and news editors and received very positive feedback via a questionnaire[1]. Based on this, we developed a set of guidelines to inform the process of summary authoring. Whilst clear about what the general nature of the target summary should be, the guidelines avoid being too prescriptive, leaving authors some freedom to include what feels intuitively correct to include in the summary for any given conversation.

3 A Method for Human Authoring of Reader Comment Summaries

To help people write overview summaries of reader comments, we have developed a 4-stage method, which is described below[2]. Summary writers are provided with an interface, which guides annotators through the 4-stage process, presenting texts in a form convenient for annotation, and collecting the annotations. The interface has been designed to be easily configurable for different languages, with versions for English, French and Italian already in issue. Key details of the methodology, guidelines and example annotations follow. Screenshots of the interfaces supporting stages 1 and 3 can be found in the Appendix.

Stage 1: Comment Labeling In this stage, annotators are shown an article in the interface, plus its comments (including the online name of the

[1]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/`.

[2]The method described here is not unlike the general method of thematic coding widely used in qualitative research, where a researcher manually assigns codes (either pre-specified and/or "discovered" as the coding process unfolds) to textual units, then groups the units by code and finally seeks to gain insights from the data so organised (Saldana, 2015). Our method differs in that: (1) our "codes" are propositional paraphrases of viewpoints expressed in comments rather than the broad thematic codes, commonly used in social science research, and (2) we aim to support an annotator in writing a summary that captures the main things people are saying as opposed to a researcher developing a thesis, though both rely on an understanding of the data that the coding and grouping process promotes.

1. Comment: *"Smart machines now collect our highway tolls, check us out at stores, take our blood pressure ..." And yet unemployment remains low.*

   Label: smart machines now carry out many jobs for us (collect tolls; checkout shopping; take blood pressure), but unemployment stays low.

2. Comment: *Not compared to the 70s, only relative to the 80s/90s.*

   Label: disagrees with 1; unemployment is not low compared to the 70's; is low relative to the 80's/90's

Figure 2: Two comments with labels (source: `www.theguardian.com/commentisfree/2016/apr/07/robots-replacing-jobs-luddites-economics-labor`).

poster, and reply-to information). Annotators are asked to write a 'label' for each comment, which is a short, free text annotation, capturing its essential content. A label should record the main "points, arguments or propositions" expressed in a comment, in effect providing a mini-summary. Two example labels are shown in Figure 2.

We do not insist on a precise notation for labels, but we advise annotators to:

1. record when a comment agrees or disagrees with something/someone
2. note grounds given in support of a position
3. note jokes, strong feeling, emotional content
4. use common keywords/abbreviations to describe similar content in different comments
5. return regularly to review/revise previous labels, when proceeding through the comments
6. make explicit any implicit content that is important to the meaning, e.g. "unemployment" in the second label of the figure (note: this process can yield labels that are longer than the original comment).

The label annotation process helps annotators to gain a good understanding of key content of the comments, whilst the labels themselves facilitate the grouping task of the next stage.

Stage 2: Label Grouping In stage 2, we ask annotators to sort through the Stage 1 labels, and to group together those which are similar or related. Annotators then provide a "Group Label" to describe the common theme of the group in terms of e.g. topic, propositions, contradicting viewpoints, humour, etc. Annotators may also split the labels in a group into "Sub-Groups" and assign a "Sub-Group Label". This exercise helps annotators to make better sense of the broad content of the comments, before writing a summary.

The annotation interface re-displays the labels created in Stage 1 in an edit window, so the annotator can cut/paste the labels (each with its comment id and poster name) into their groups, add Group Labels, and so on. Here, annotators work mainly with the label text, but can refer to the source comment text (shown in context in the comment stream) if they so wish. When the annotator feels they have sorted and characterised the data sufficiently, they can proceed to stage 3.

Stage 3: Summary Generation Annotators write summaries based on their Label-Grouping analysis. The interface (Figure 5) displays the Grouping annotation from Stage 2, alongside a text box where the summary is written in two phases. Annotators first write an 'unconstrained summary', with no word-length requirement, and then (with the first summary still visible) write a 'constrained-length summary' of 150–250 words.

Further analysis may take place as a person decides on what sentences to include in the summary. For example, an annotator may:

- develop a group label, e.g. producing a polished or complete sentence;
- carry out further abstraction over the groups, e.g. using a new high-level statement to summarise content from two separate groups;
- exemplify, clarify or provide grounds for a summary sentence, using details from labels or comments within a group, etc.

We encourage the use of phrases such as "many/several/few comments said...", "opinion was divided on...", "the consensus was...", etc, to quantify the proportion of comments/posters addressing various topics/issues, and the strength/polarisation of opinion/feeling on different issues.

Stage 4: Back-Linking In this stage, annotators link sentences of the constrained-length summary back to the groups (or sub-groups) that informed their creation. Such links imply that at least some of the labels in a group (or sub-group) played a part supporting the sentence. The interface displays the summary sentences alongside the Label Grouping from Stage 2, allowing the annotator to select a sentence and a group (or sub-group — the more specific correct option is preferred) to assert a link between them, until all links have been added. Note that while back-links are to groups

of *labels*, the labels have associated comment ids, so indirectly summary sentences are linked back to the source comments that support them. This last stage goes beyond the summary creation process, but captures information valuable for system development and evaluation.

4 Corpus Creation

4.1 Annotators and training

We recruited 15 annotators to carry out the summary writing task. They included: final year journalism students, graduates with expertise in language and writing, and academics. The majority of annotators were native English speakers; all had excellent skills in written English. We provided a training session taking 1.5-2 hours for all annotators. This included an introduction to our guidelines for writing summaries.

4.2 Source Data

From an initial collection of 3,362 *Guardian* news articles published in June-July 2014 and associated comment sets, we selected a small subset for use in the summary corpus. Articles were drawn from the *Guardian*-designated topic-domains: politics, sport, health, environment, business, Scotland-news and science. Table 1 shows the summary statistics for the 18 selected sets of source texts (articles and comments). The average article length is 772 words. The comment sets ranged in size from 100 to 1,076 comments. For the annotation task, we selected a subset of each full comment set, by first ordering threads into chronological order (i.e. oldest first), and then selecting the first 100 comments. If the thread containing the 100^{th} comment had further comments, we continued including comments until the last comment in that thread. This produced a collection of reduced comment sets totalling 87,559 words in 1,845 comments. Reduced summary comment sets vary in length from 2,384 words to 8,663 words.

5 Results and Analysis

The SENSEI Social Media Corpus, comprising the full text of the original *Guardian* articles and reader comments as well as all annotations generated in the four stage summary writing method described in Section 3 above – comment labels, groups, summaries and backlinks – is freely available at: `nlp.shef.ac.uk/sensei/`.

5.1 Overview of Corpus Annotations

There were 18 articles and comment sets, of which 15 were double annotated and 3 were triple annotated, giving a total of 39 sets of complete annotations. Annotators took 3.5-6 hours to complete the task for an article and comment set.

Table 2 shows a summary of corpus annotations counts. The corpus includes 3,879 *comment labels*, an average of 99.46 per annotation set (av. 99.46/AS). There are, in total, 329 *group annotations* (av. 8.44/AS) and 218 *subgroups* (av. 5.59/AS). Each of the 547 groups/subgroups has a short *group label* to characterise its content. Such labels range from keywords ("midges", "UK climate", "fining directors", "Air conditioning/fans") to full propositions/questions ("Not fair that SE gets the investment", "Why use the fine on wifi?"). Each of the 39 annotation sets has two summaries, of which the *unconstrained summaries* have average length 321.41 words, and the *constrained summaries*, 237.74 (a 26% decrease). Each summary sentence is back-linked to one or more groups comment labels that informed it.

5.2 Observations

Variation in Grouping There is considerable variation between annotators in use of the option to group/sub-group comment labels. Whilst the average of groups per annotation set was 9.0, for the annotator who grouped the least this was 4.0, and the maximum average 14.5. For sub-groups, the average per annotation set was 5.0. 14 of 15 annotators used the sub-group option in at least one annotation set, and only 5 of the 39 sets included no sub-groups. A closer look shows a divide between annotators who use sub-groups quite frequently (7 having an average of $\geq$6.5/AS) and those who do not (with av. $\leq$2/AS).

Other variations in annotator style include the fact that around a third of them did most of their grouping at the sub-group level (4 of the 6 who frequently used subgroups were amongst those having the lowest average number of groups). Also, whilst a fifth of annotators preferred to use mainly a single level of grouping (i.e. had a high average of groups, and a low average of sub-groups, per annotation set), another fifth of annotators liked to create both a high number of groups and of sub-groups, i.e. used a more fine-grained analysis.

We also investigated whether the word-length of a comment set influenced the number of

	Total	Min	Max	Mean
Article and Comment Sets(number)	18	-	-	-
Article, word length	13,898	415	2,021	772.11
Full Comment Set, total word length	318,618	4,918	37,543	17,701
Full Comment Set, total comments	6,968	100	1,076	387.11
Reduced Comment Set (number)	18	-	-	-
Reduced Comment Set, total comments	1,845	100	109	102.5
Reduced Comment Set, total word length	87,559	2,384	8,663	4,864.39
Reduced Comment Set, single comment word length	-	1	547	47.46

Table 1: Summary Statistics for Corpus Source Texts

	Total	Min	Max	Mean
Annotated Comment Set (number)	18	-	-	-
Completed Annotation Sets (number)	39	-	-	-
Stage 1 Labels (number)	3,879	69	109	99.46
Length of Unconstrained Summaries (words)	12,535	131	664	321.41
Length of Constrained Summaries (words)	9,272	152	249	237.74
Number of Groups / Group Labels	329	4	17	8.44
Number of Sub-Groups / Sub-Group Labels	218	0	15	5.59
Number of Labels in Groups	4,050	1	84	12.31
Number of Labels in Sub-groups	1,435	1	27	6.58
Note: Total count, min, max and mean are drawn from across the full set of corpus annotations				

Table 2: Annotation Statistics

groups/subgroups created by the annotators, but surprisingly, there was no obvious correlation.

Reader Comment Summaries We carried out a preliminary qualitative analysis to establish the character of the summaries produced, which shows that they are in general all coherent and grammatical, and that the majority of summary sentences characterise views on issues. Some observations on summary content follow:

1. All summaries contain sentences reporting different *views* on issues. Figure 2 shows two typical summaries, which describe a range of views on two main issues: "whether or not citizens can cope with reductions in bin collection" (Summary 1), and "whether or not new taxes on the rich should be introduced to pay for the NHS" (Summary 2).

2. Summaries frequently indicate points of contention or counter arguments, e.g. sentences *(S2)* and *(S5)* of Summary 2.

3. Summaries often provide examples of the reasons people gave in support of a viewpoint: e.g. *(S2)* of Summary 1 explains that people thought a reduced bin collection would attract vermin because the bins will overflow with rubbish.

4. Annotators often indicate the proportion/amount of comment addressing a particular topic/issue or supporting a particular viewpoint, e.g. see *(S6)* of Summary 2; *(S3)* of Summary 1.

5. While the majority of annotators abstracted across groups of comments to describe views on issues, there were a few outliers who did not. For example, for an article about a heatwave in the UK, the two annotators grouped the same 8 comments, but summarised the content very differently. Annotator 1 generalised over the comments: "A small group of comments discussed how the heat brings about the nuisance of midges and how to deal with them". Annotator 2 listed the points made in successive comments: "One person said how midges were a problem in this weather, another said they should shut the windows or get a screen. One person told an anecdote about the use of a citronella candle . . . another said they were surprised the candle worked as they had been severely bitten after using citronella oil".

6. Very few summary sentences describe a discussion topic without indicating views on it (e.g. "Many comments discuss the disposal of fat").

Analysis revealed that summaries also include examples of: *Background* about, e.g., an event,

Summary 1

(S1) Opinions throughout the comments were divided regarding whether residents could cope with Bury's decision to collect grey household bins every three weeks rather than every two, and the impact this could have on households and the environment. *(S2)* Some argued how the reduction in bin collection would attract vermin as bins overflow with rubbish, while others gave suggestions of how waste could be reduced. *(S3)* The largest group of commenters reflected on how successful (or not) their specific bin collection scheme was at reducing waste and increasing recycling. *(S4)* Throughout the comments there appeared to be some confusion on what waste could be recycled in the grey household bin in Bury. *(S5)* It also appeared unclear if Bury currently provides a food waste bin and if not one commenter suggested that the borough should provide one in the effort to reduce grey bin waste. *(S6)* A large number of comments suggested how residents could reduce the amount of waste going into the grey household bin by improving their recycling behaviour. *(S7)* This led to a deeper discussion regarding the pros and cons of reusable and disposable nappies...

Summary 2

(S1) The majority of people agreed that businesses and the rich should pay more tax to fund the NHS, rather than those on low incomes. *(S2)* Some said income tax should be raised for the highest earners and others suggested a 'mansion tax'. *(S3)* Some commenters suggested that the top one percent of earners should pay up to 95 in income tax. *(S4)* Although, there was a debate as to how 'rich' can be defined fairly. *(S5)* Other commenters pointed out that raising taxes would damage the economy and drive the most talented minds and business to different countries with lower taxes. *(S6)* A large proportion of commenters said the government should do more to tackle tax evasion and avoidance by big businesses and the rich. *(S7)* But some said the extent of tax evasion was exaggerated by the press. *(S8)* A strong number of people criticised the coalition for cutting taxes for the rich and placing the burden on lower-paid workers. *(S9)* They said that income tax has been cut for the very rich, while benefits have been slashed and VAT has increased, making life for low-paid workers more difficult. *(S10)* Many criticised the Liberal Democrats for going into a coalition with the Conservatives and failing to keep promises. *(S11)* Many said they had failed to curb Tory excesses and had abandoned their core principles and pledges. *(S12)* A small minority said that the NHS is too expensive and needs reform.

Figure 3: Two human authored summaries of comment sets. These summaries and the source articles and comments are in the SENSEI Corpus available at: `nlp.shef.ac.uk/sensei`.

practice or person, to clarify an aspect of the debate, e.g. see *(S5)* of Summary 1, *Humour*; *Feelings* and *Complaints*, about e.g. commenters and reporters.

5.3 Similarity of Summary Content

We investigated the extent to which summaries of the same set of comments by different annotators have the same summary content, by performing a content comparison assessment on 10 randomly selected summary pairs, using a method similar to the manual evaluation method of DUC 2001 (Lin and Hovy, 2002).

Given summaries A and B, for each sentence s in A, a subject judges the extent to which the meaning of s is evidenced (anywhere) in B, assigning a score on a 5-point scale (5=all meaning evidenced; 1=none is). Any score above 1 requires evidence of *common propositional content* (i.e., a common entity reference alone would not suffice). After A is compared to B, B is compared to A.

Comparison of the 10 random summary pairs required 300 sentence judgements, which were each done twice by two judges and averaged. In these results, 17% of summary sentences received a score of 5 (indicating all meaning evidenced) and 40% a score between 3 and 4.5 (suggesting some or most of their meaning was evidenced). Only 15% of sentences received a score of 1.

Looking at the content overlap per individual summary pair (by averaging the sentence overlap scores for that pair), we find values for the 10 pairs that range from 2.56 up to 3.65 (with overall average 3.06). Scores may be affected by the length of comment sets (as longer sets give more scope for variation and complexity), and we observe that the two lowest scores are for long comment sets.

We assessed the agreement between judges on this task, by comparing their scores for each sentence. Scores differ by 0 in 46% of cases, and by 1 in 33%, giving a combined 79% with 'near agreement'. Scores differ by >2 in only 6% of cases. These results suggest that average sentence similarity is a reliable measure of summary overlap.

6 Related Work

Creating abstractive reference summaries of extended dialogues is hard. A more common approach involves humans assessing source units (e.g., comments in comment streams, turns in email exchanges) based on their perceived importance (aka "salience") for inclusion in an end summary. See, e.g., Khabiri et al.'s (2011) work on comments on YouTube videos; Murray and Carenini's (2008) work on summarizing email discussions. The result is a "gold standard" set of units, each with a value based on multiple human annotations. A system generated extractive summary is then scored against this gold standard. The underlying assumption is that a good summary of length n is one that has a high score when compared against the top-ranked n gold standard units.

Such an approach is straightforward and provides useful feedback for extractive summarization systems. While the gold standard is extractive, the selected content may have an abstractive flavour if annotators are instructed to favour "meta-level" source units that contain overview content. But the comment domain has few obvious examples of meta-level sentences; explicit references to the issues under discussion are few, as are reflective comments that sum up a preceding series of comments. Moreover, extractive approaches to writing comment summaries will almost certainly fall short of indicating aggregation over views and opinion. In sum, this is not an ideal approach to creating reference summaries from comment.

A more abstractive approach to writing summaries of multi-party conversations was used in the creation of the AMI corpus annotations, based on 100 hours of recorded meetings dialogues (Carletta et al., 2006). There are some similarities and differences between the AMI approach and our own. First, AMI summary writers first completed a topic segmentation task to prepare them for the task of writing a summary. While segmentation might appear to resemble our *grouping* stage, these are very different tasks. Key differences are that segmentation was carried on AMI dialogues using a pre-specified list of topic descriptions. This would be difficult to provide for comment summary writers, since we cannot predict everything the comments will talk about. Secondly, the AMI abstractive summaries are linked to dialogue acts (DAs) in their manual extractive summaries (a link is made if a DA is judged to "support" a sentence in the abstractive summary). Similar to our back-links, their links provide indices from the abstractive summary to source text units. However, our back-links are from a summary sentence to *groups* of comment labels that the summary author has judged to have informed his sentence. Finally, the AMI abstractive summaries comprise an overview summary of the meeting, and list "decisions", "problems/issues" and "actions". However, while a very small number of non-scenario corpus summaries included reports of alternative views in a meeting (e.g. on which film to choose for a film club), the AMI scenario summaries include very few examples of differences in opinion.

Misra et al. (2015) have created manual summaries of short dialogue sequences, extracted from different conversations on similar issues on debating websites. They then collected summaries together, and applied the Pyramid method (Nenkova et al., 2007) to identify common, central propositions, which, they describe as "abstract objects" that represent facets of an argument on an issue, e.g. gay marriage. Indeed the task of identifying central propositions across multiple conversations is a key aim in their work and one they point out is central to others working in argumentation mining. They use the Pyramid annotations to provide indices from the central proposition to the summary and underlying comment, with a view to learning how to recognize similar argument facets automatically. Note their task differs from ours in that we aim to generate a summary of a single reader comment conversation, while they aim to identify (and then possibly summarize) all facets of a single argument, gleaned from multiple distinct conversations.

Barker and Gaizauskas (2016) elaborate the issue-viewpoint-evidence framework introduced in Section 2.1 above and show how an argument graph representing an analysis in this framework may be created for a set of comments. They show how the content in a single reference summary, created using the informal label and group method described above, corresponds closely to a subgraph in the more formally specified argument graph for the article and comment set.

7 Concluding Remarks and Future Work

We have presented a proposal for a form of informative summary that aims to capture the key content of multi-party, argument-oriented conversations, such as those found in reader comment. We have developed a method to help humans author such summaries, and used it to build a corpus of reader comment multiply annotated with summaries and other information. We believe the method of labeling and grouping has wide application, i.e. in creating reference summaries of complex, multi-party dialogues in other domains.

The summaries produced correspond closely to the target specification given in Sec. 2.2, and exhibit a high degree of consistency, as shown by the content similarity assessment of Sec. 5.3. Informal feedback from media professionals (at the *Guardian* and elsewhere) suggests that the summaries are viewed very positively as a summary of comments in themselves, and as a target for what an automated system might deliver online.

Our summary corpus has already proved useful in providing insights for system development, and for training and evaluation. We have used group annotations to evaluate a clustering algorithm (Aker et al., 2016a); used back-links to inform the training of a cluster labeling algorithm (Aker et al., 2016b); used the summaries as references in evaluating system outputs (with ROUGE as metric), and to inform human assessors in a task-based system evaluation (Barker et al., 2016).

Even so, there are limitations to the work done which give pointers to further work. The current corpus is limited in size, and would ideally contain annotations for more comment sets, with more annotations per set. One possibility is to break the summary creation method into smaller tasks suitable for crowd-sourcing. Another issue is scalability: annotators can write summaries for $\sim$100 comments, but this is time-consuming and taxing, casting doubt on whether the method could scale to 1000 comments. Results from a pilot suggest annotators find it much easier to work on sets of 30–50 comments, so we are investigating how annotations for smaller subsets of a comment set might be merged into a single annotation.

Many of our annotators found the option to have groups *and* sub-groups useful, but this feature presents problems for some practical uses of the annotations, such as evaluation of some clustering methods. Hence, we have investigated methods to flatten the group-subgroup structure into one level, including the following two methods: (1) simple flattening, where all sub-groups merge into their parent groups (but this loses much of the analysis of some annotators), and (2) promoting subgroups to full group status (which has proved useful for generating useful group labels). More research is needed to establish the most effective flattening to best capture the consensus between annotators.

Finally, there is the open question of how to automatically evaluate system-generated summaries against the reference summaries proposed here. In particular, is ROUGE (Lin, 2004), the most widely used metric for automatic summary evaluation, an appropriate metric for use in this context? ROUGE, which calculates n-gram overlap between system and reference summaries, may not deal well with the abstractive nature of our summaries, and in particular with statements quantifying the distribution of support for various viewpoints. Its utility needs to be established by correlating it with human judgements on system output quality. If it cannot be validated, the challenge arises to develop a metric better suited to this evaluation need.

Acknowledgments

The authors would like to thank the European Commission for supporting this work, carried out as part of the FP7 SENSEI project, grant reference: FP7-ICT-610916. We would also like to thank all the annotators without whose work the SENSEI corpus would not have been created, Jonathan Foster for his help in recruiting annotators and *The Guardian* for allowing us access to their materials. Finally, thanks to our anonymous reviewers whose comments and suggestions have helped us to improve the paper.

References

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

Ahmet Aker, Monica Paramita, Emina Kurtic, Adam Funk, Emma Barker, Mark Hepple, and Rob Gaizauskas. 2016b. Automatic label generation for news comment clusters. In *Proceedings of the 9th International Conference on Natural Language Generation Conference (INLG)*, Edinburgh, UK.

Emma Barker and Robert Gaizauskas. 2016. Summarizing multi-party argumentative conversations in reader comment on news. In *Proceedings of the 3rd Workshop on Argument Mining*, Berlin.

Emma Barker, Monica Paramita, Adam Funk, Emina Kurtic, Ahmet Aker, Jonathan Foster, Mark Hepple, and Robert Gaizauskas. 2016. What's the issue here?: Task-based evaluation of reader comment summarization systems. In *Proceedings of LREC 2016*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39, Berlin, Heidelberg. Springer-Verlag.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proc. of the First Workshop on Argumentation Mining*, pages 39–48.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 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.

Elham Khabiri, James Caverlee, and Chiao-Fang Hsu. 2011. Summarizing user-contributed comments. In *Proceedings of The Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, pages 534–537, Barcelona.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, AS '02, pages 45–51, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out*, jul.

Clare Llewellyn, Claire Grover, and Jon Oberlander. 2014. Summarizing newspaper comments. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 265–274, New York, NY, USA. ACM.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado, May–June. Association for Computational Linguistics.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 773–782, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4, May.

Johnny Saldana. 2015. *The Coding Manual for Qualitative Researchers*. Sage Publications Ltd, 3 edition.

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

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proc. of the SIGDIAL 2015 Conference*, pages 217–226. Association for Computational Linguistics.

Appendix

Comment Number	User Name	Comment Communication	Comment	Comment Label
1	lindalusardi	**1**	what about having a word with them about ticket prices aswell?	NR - ticket prices (implies too high)
2	Cynic24	**2** --> 1	That would be fairly pointless, given that Network Rail don't operate passenger trains!	NR do not set fares / operate trains
3	Cynic24	**3** --> 1	The fact that the first post is getting recommends shows that there are some clueless people out there! If you are going to criticise the railways (I do so fairly often), then at least read up on how they actually work! Criticising from from a position of complete ignorance won't gain you any credibility! Passenger trains are not operated by Network Rail (their remit is solely to maintain the railway infrastructure), therefore the ticket prices are nothing to do with them!	NR do not set fares / operate trains
4	lindalusardi	**4** --> 2	silly me, of course I forgot that network rail provide their services to rail operators completely free and have nothing whatsoever to do with increasing ticket prices d'oh	NR do not set fares - accepted
5	martin77	**5** --> 3	OK so the lady made a mistake , calm down for Heavens sake.	non-topic; calm down
6	Craig Axon	**6** --> 2	Average ticket prices are set by the government. The train company then as a leeway to increase or decrease the price from the average set. Ticket prices are determined not by distance, as such, but by the number of people travelling from one particular point to another and so forth. There's more to it but that's the gist of it. Also, most the money made by fares is invested back into the railway infrastructure, so it's a cycle. Only a small amount per £ spent on a ticket goes to train company profits. Most goes to network rail, fuel, train maintenance, staff costs and so forth. It must also be noted that network rail is a not for profit company and many profits either get invested back into the railway or given to the government to find major projects like thameslink and crossrail. The railway is not very transparent which is the problem why most people vent their rage on train companies and network rail.	Ticket prices set by government; NR not-for-profit; system not transparent

Figure 4: Stage 1 interface. The first 4 columns are created automatically from the source reader comments. The last column is a label supplied by the annotator.

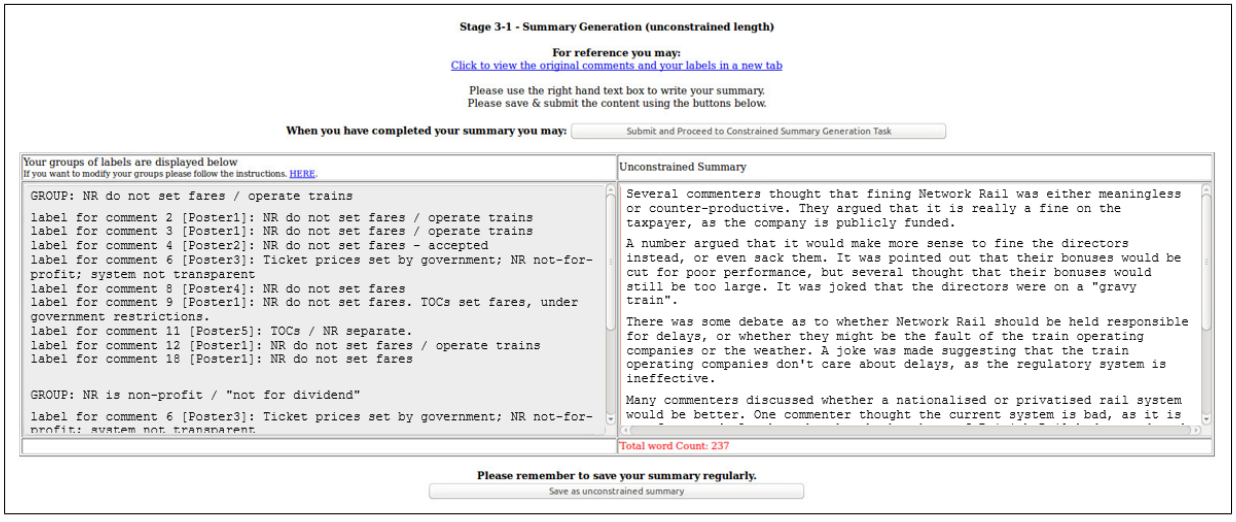

Figure 5: Stage 3 interface. Grouping annotations collected in Stage 2 are shown in the left frame. The summary is authored in the right frame.