{SENSEI}

D1.2 – Report on Use Case Design and User Requirements

Document Number	D1.2
Document Title	Report on Use Case Design and User Requirements
Version	3.2
Status	Final
Work Package	WP1
Deliverable Type	Report
Contractual Date of Delivery	31.10.2014
Actual Date of Delivery	31.10.2014
Responsible Unit	UNITN
Keyword List	Use Case, Evaluation, User Requirements
Dissemination level	PU

Editors

Morena Danieli (UNITN)

Rob Gaizauskas (USFD)

Contributors

Frédéric Béchet	(AMU)
Emma Barker	(USFD)
Cosima Caramia	(TP)
Morena Danieli	(UNITN)
Benoit Favre	(AMU)
Rob Gaizauskas	(USFD)
Mark Hepple	(USFD)
Letizia Molinari	(TP)
Adele Palumbo	(TP)
Monica Paramita	(USFD)
Massimo Poesio	(UESSEX)
Giuseppe Riccardi	(UNITN)
Danilo Rizzo	(TP)

SENSEI Coordinator

Prof. Giuseppe Riccardi

Department of Information Engineering and Computer Science

University of Trento, Italy

riccardi@disi.unitn.it

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	29/07/2014	Draft	Morena Danieli (UNITN)	Table of Content and outline (who does what)
0.4	01/08/2014	Draft	Morena Danieli (UNITN)	Added Names of Contributors
0.5	08/07/2014	Draft	Morena Danieli (UNITN) Letizia Molinari (TP)	Inserted text in par 2.2 and related references
0.6	08/08/2014	Draft	Morena Danieli (UNITN)	Section 2.3, initial specification of speech conversation summary use case.
0.7	08/28/2014	Draft	Benoit Favre (AMU)	Formatting of the document; Section 3.2, extrinsic evaluation
0.8	08/29/2014	Draft	Morena Danieli (UNITN) Letizia Molinari (TP)	Section 2.3 and 1.2.1
0.9	08/29/2014	Draft	Benoit Favre (AMU)	Section 3.1 intrinsic evaluation and 3.4 baselines
1.0	09/01/2014	Draft	Morena Danieli (UNITN) Letizia Molinari, Cosima Caramia, Adele Palumbo (TP)	Appendix - ACOF and notes; Removed the paragraph on insight from annotation because it will be part of the accompanying document of the annotated database (Milestone 1).
1.2	09/09/2014	Draft	Morena Danieli (UNITN)	Overview paragraph, and revision of the initial summary
1.3	09/10/2104	Draft	Emma Barker (USFD)	Inserted revised Social Media Use Cases, Appendix containing questionnaires and redrafted some parts of the methodology section.
1.4	09/12/2014	Draft	Morena Danieli (UNITN)	Moved on Dropbox, some editing; Table of Contents.
1.5	09/22/2014	Draft	Benoit Favre (AMU)	Update metrics section
1.6	09/24/2014	Draft	Emma Barker and Rob Gaizauskas (USFD)	Updated Social Media Use Cases and Appendix containing questionnaires. Added

				sections on intrinsic, extrinsic and insight-oriented evaluation for the Social Media Use Case
1.7	09/24/2014	Draft	Morena Danieli (UNITN) Letizia Molinari, Adele Palumbo, Cosima Caramia (TP)	General Editing, inserted Coreference metrics, Synopses Guidelines, Completed the Executive Summary and Overview sections. Still to be done: report focus groups data in Appendix A, and Conclusion.
1.8	09/24/2014	Draft	Morena Danieli (UNITN)	Update metrics section, restructured the evaluation section, Completed Appendix A
1.9	09/28/2014	Draft	Morena Danieli (UNITN)	Drafted Conclusion, and arranged Annotation Guidelines Appendix
1.11	10/04/2104	Draft	Rob Gaizauskas, Emma Barker (USFD	Revised section on extrinsic evaluation – introduction and sub-section on social media
2.0	10/06/2014	Draft	Benoit Favre (AMU)	Update speech evaluation
2.0-GR	10/06/2014	Draft	Giuseppe Riccardi (UNITN)	General review, comments, executive summary rewritten
2.1	10/06/2014	Draft	Monica Paramita, Emma Barker, Rob Gaizauskas (USFD)	Updated social media use cases and Appendix D
2.2	10/08/2014	Draft	Elisa Chiarani (UNITN)	Quality check completed
2.3	10/08/2014	Draft	Morena Danieli (UNITN)	Added keywords, revised on the basis of Elisa's comments
2.4	10/8/2014	Draft	Hugo Zaragoza (WEBSAYS)	Social Media Use cases review. Overall Quality check
2.5	10/14/2014	Draft	Morena Danieli (UNITN)	Final review
2.6	10/17/2014	Draft	Emma Barker, Monica Paramita, Mark Hepple, Rob Gaizauskas	Added results from Social Media Use Case Questionnaire
3.0	10/20/2014	Final	Elisa Chiarani, Giuseppe Riccardi (UNITN)	Quality check and finalisation for Review

3.1	30/10/2014	Draft	Emma Barker, Rob Gaizauskas (USFD)	Revised sections on intrinsic, extrinsic and insight-oriented evaluation for the Social Media Use Case. Added Appendix F.
3.2	30/10/2014	Final	Morena Danieli (UNITN)	Final review

Executive summary

SENSEI deliverable D1.2 presents the research that led to the final selection of the speech and social media use cases presented in D1.1, and the evaluation model that will be used throughout the project.

The selection process has included interaction with groups of potential users from both speech and social media domains. The user groups involved in the social media use case are journalists, comment posters and comment readers; the user groups involved in the speech use case are call centre professionals working in the Quality Assurance division of a call centree company. The interactions with both groups of potential users aimed at identifying the user needs and requirements that could be met by SENSEI objectives.

For the two scenarios the findings of the qualitative investigation showed that SENSEI technologies can improve the quality and effectiveness of the complex processes of call centre agent monitoring (speech), and provide users with new tools based on summarization techniques (social media). On the basis of those findings, for the speech scenario two target use cases were selected. The first use scenario aims to provide automatic generation of Agent Conversation Observation Forms (ACOFs), i.e. call surveys focused on rating agent behavior according to a set of desired conversational behaviours. The second use case is focused on the generation of call centre conversation summaries and aggregation of calls on the basis of topic identification. For the social media scenario the six use cases have been re-drafted and reformulated for being presented to the users.

The second part of the document presents the SENSEI evaluation model, and the related evaluation scenarios. We have designed a tripartite evaluation model that includes three kinds of metrics: technology-oriented (intrinsic), task-oriented (extrinsic), and insight-oriented metrics.

Table of Content

Introduction

This deliverable reports the results of the work done during months M6 - M12 in the SENSEI Work Package 1. It builds upon the results described in D1.1. The document is structured as follows. The first part of the document (Sections 1-3) reports the speech and social media use cases and the methods applied to select, prioritize, and refine them. For both scenarios the methodology applied included qualitative research tools such as interviews, questionnaires, and focus groups with users. For both application domains, the materials used in this process are reported in the Appendices at the bottom of the present document. The refinement of the speech and social media use cases allowed enabled the identification of the SENSEI products, which will include the automatic generation of surveys and summaries for call centre conversations, and summarization/analytics of news articles and comments. The second part of the document illustrates the SENSEI evaluation methodology and the evaluation scenarios. This content is organized into a tripartite model of evaluation that includes extrinsic, intrinsic, and insight-oriented evaluation metrics. The definition of the SENSEI use cases also includes the annotation of initial corpora of call centre conversations, news articles and comments.

1. The SENSEI Use Cases

In this section we present the refined versions of the use cases previously described in SENSEI Deliverable D1.1 Preliminary Version of Use Case Design. We begin by describing the methodology we used to refine the use cases, particularly the methods we used to elicit user requirements and user feedback on the imagined use cases proposed in D1.1. Then each of the use cases selected for refinement is described in a dedicated sub-section, with details about scenarios of use, input data, and the expected results. Appendices at the end of this document contain details of data actually collected by focus groups, interviews with end users, and so on.

1.1. Methods for Use Case Refinement

In this section we describe the procedures we followed for gathering user requirements or soliciting feedback/comment from users on proposed use cases. In both domains of application we have been basing the final stage of use case development by consulting potential users, and refining and modifying the original use cases according to their feedbacks.

For the speech scenario we run two focus groups with two categories of final users (the coordinator of the front office supervisors, and two front office agents' supervisors) and an interview with the Quality Assurance Director.

Compared with survey methods focus groups are expected to provide more insight and to be more naturalistic (Berg & Lune 2004; Grudens-Schuck, Allen & Larson, 2004). The speech use case focus groups were run with the features summarized in Table 1 below.

Table 1 - Elements of the speech use case focus groups

Elements of the speech use case focus groups	
Format	Group session
Size	4
Length	1.5 hours
Number of sessions	2
Participants	Selected, by invitation, with similar characteristics; 3 out of 4 were invited twice.
Forms of data	Wordings and issues emerging from conversation
Additional materials	ACOFs, Summary/Synopsis Guide, and sample of summaries
Data collection	Moderator notes
Uses interview guide	Defined in the first session, modified on the basis of feedback and focused on summaries in the second session
Format of reporting	Selected quotations and analysis of repeated issues

Two group sessions were held in TP premises on April, 2nd (Taranto, Italy) and July, 16th (Rome, Italy). We had four participants in the first session (two managers, male and female, and two front office supervisors, female), as well as in the second one (three front office supervisors, and one manager, all female).

As we will explain in details in the next paragraphs, using this qualitative research methodology allowed us to define the SENSEI Agent Observation Form, and to evaluate the interest of potential users into possible use of conversation summaries by QA professionals. While in the first case, what SENSEI can provide is something that can impact on an already existing activity, i.e. the generation of call surveys, in the second case the participants were asked to hypothesize if their work could benefit from the availability of automatically generated conversation summaries.

From the focus group discussion it emerged that potential users have positive expectations concerning possible improvements deriving from SENSEI results, both in terms of reduction of time-to-completion in the activity related with agent monitoring, and in the opportunity of focusing QA professional attention on problematic calls. Actually in a call centre there are lot of data continuously tracked and measured, but from the focus group discussion we could learn that the issues related with agent motivation, customer satisfaction, and identification of agent training needs can be effectively monitored only on the basis of competent human listening. Since it was showed in many studies that factors including motivation and organizational identification have a positive influence on job efficiency and effectiveness, we chose to focus on the SENSEI use cases that we envisage as more promising for improving the human monitoring of call centre agents.

For the Social Media we employed two main routes for gathering data and feedback from users on our characterization of the user groups and the use cases (Full details are provided below in Section 3.2):

1. **Questionnaires** – these are presented via a web interface, and competed remotely, sometimes with the support of a phone call to support the form filling process.

2. **Informal Discussion with users.**

2. Speech Use Cases

In this section we describe the details of the scenarios of the two speech use cases that have been selected, and the process we implemented for identifying possible uses of the SENSEI results.

2.1. Automatic Generation of Call Surveys

The automatic generation of call surveys is the fourth use case of the speech domain that was introduced in D1.1. The SENSEI Consortium selected this use case for further studying it in view of implementing its requirements.

The goal we aim to achieve is to provide the Quality Assurance Supervisors of call centres with automatically generated Agent Conversation Observation Forms (ACOFs henceforth) for each inbound call.

We have investigated the generation process of ACOFs in a real call centre environment that is the TP site in Taranto by interviewing the Front Office Supervisor manager, by observing the job of Front Office Supervisors, and by observing the Agents while they replied to inbound calls of a customer care service. In the following subsections we illustrate the results of such investigation (2.2.1.1), we introduce and comment the ACOF that is being used for annotating the calls of the DECODA and LUNA corpora (2.2.1.2), and finally we describe how the automatic generated ACOFs can be used by the call centre Front Office Supervisors.

2.1.1. The generation process of ACOFs

ACOFs are checklists that Quality Assurance Supervisors fill in while surveying call centre agents. While most of the ACOFs aim to evaluate general conversational skills, some items can depend on the particular campaign. For example, the client company may require that the agent speech comply with the company requirements for welcoming, closing the call, naming the company product, etc. The ACOF is acknowledged by the Quality Manager, who can propose changes related with the quality assurance policy of the call centre. For each inbound or outbound campaign the agents receive a focused training, and they also know which ACOF version their supervisors will use.

The ACOF checklists are usually arranged into five behavioral classes, i.e. the call opening and closing, the management of errors, client objections, and selling abilities (if relevant). Each item has an associated score, which is automatically generated on the basis of the Agent Supervisor evaluation.

The ACOFs are used for monitoring of the qualitative trend of the call centre campaigns. The qualitative data that can be extracted by the surveys are correlated with quantitative data tracked by the back office, for example the number of new contracts in a selling campaign. Qualitative and quantitative correlations are assessed by applying nonparametric measure of statistical dependence, such as Spearman's rank correlation coefficient (Zar, 1972).

Since the ACOFs aim to assess in the agent performance the presence of a required behavior, it is likely that a single call listening does not provide the Quality Assurance Supervisor with behavior instances sufficient for the task. We will come back to this issue in the evaluation part of this Deliverable.

2.1.2. Agent Behavior in Focus

On the basis of the analysis of currently used ACOFs in the TP call centres, we abstracted a set of agent behavior that is independent from the particular campaign. The set is the following:

1. Communication Skills
2. Call Opening and Closing
3. Problem Solving
4. Proactivity

Let us illustrate them in more details.

Communication Skills and Call Opening and Closing refer to the ability of the agent to communicate effectively, where the effectiveness is evaluated on the basis of the agent politeness and her/his capability to understand the Client. As for Call Opening and Closing, the Quality Assurance supervisors also evaluate if the agents comply with scripts they receive for opening and closing the call (initial and final greetings, presentation, identification of the client, etc…). The agents are trained in view of being able to deal with clients that either need to solve a problem (customer care), or that are called by agents in outbound campaigns. In both cases the client may be worried or frustrated. In addition, while the client knows her/his problem, s/he does not necessarily explain the problem clearly or with technical wordings. The agents are trained not only for being polite, but also for being able to adequate their speaking style to the speaking style of the client. Moreover, they receive specific training for increasing their listening abilities, for example they are required not to interrupt the client when s/he explains the problem, even if s/he complains (Cameron, 2000; Ganguly, 2009). For assessing the agent's communication skills, Quality Assurance supervisors rely not only on the speech content and wordings of the agent, but also on tone of voice (Friginal, 2008, 2009; Clark et al. 2012) and perceived empathy (Grougiou 2004).

Problem Solving and Proactivity refers to the ability of the agents in understanding the client requests, and identifying possible solutions, by possibly anticipating further issues. In addition, the Quality Assurance policies require that the agents are trained in order not only to fix problems and find solutions, but also to increase value for the buyer, because high problem solving capabilities of the agents have been shown to correlate significantly with high rate of customer satisfaction, trust, affective commitment, and customer loyalty (Jaiswal 2008), although the latter is fully mediated by relationship quality (van der Aa 2013).

2.1.3. The SENSEI ACOF

All the versions of inbound and outbound ACOFs currently adopted for agent monitoring by TP include questions about communication skills, problem solving, and proactivity. However, none of them can be directly used for SENSEI purposes, due to the fact that all of them refer to evaluation templates modified for complying with the requirements of TP client companies. That is why we have defined a SENSEI ACOF template that, while abstracting from the idiosyncrasies of application domains, focuses on the above mentioned features of agent behavior. The SENSEI ACOF is reported in the Table 2 below. More comments on ACOF items can be found in the Appendix B.

Table 2 – SENSEI ACOF

	YES	NO	NA	Notes
Call Opening				
Agent complies with call opening protocol				
Problem Identification				
Agent's listening is focused on Client's problem and his/her questions are relevant				
Problem Fixing				
Agent is competent with respect to the product and applications				
Agent explanations are clear, complete and simple				
Agent applies the correct protocol				
Increasing Value				
Agent can manage Client's objections, reassure him/her, and takes in great consideration Client's satisfaction				
Summary, Confirmation, and Call Closing				
Agent is self-confident				
Agent speech is positively connotated				
Agent complies with the call closing protocol				

Communication Skills				
Agent is polite but purposeful with the Client				
While being professional, Agent is able to adequate his/her speech to the Client's speaking style				
Time management: Agent manage the possible waiting times by explaining their rationale				
Listening capability				

As we can observe, the SENSEI ACOF includes questions that address the four classes of agent behavior described in section 2.2.1.2. As we said above, each item of the check list is associated with a score. The Quality Assurance supervisor may insert comments (brief sentences) that motivate his/her choice.

2.1.4. Speech use case ACOF: Scenario of use

The Quality Assurance supervisor's task is to listen to a given number of calls managed by an agent with the goal of filling the ACOF checklist. As discussed in D1.1, the goal of the listening is related not only with the monitoring in view of quality assurance, but also with identifying problematic calls that may suggest the need of further training. The call centre agents know both that their job may be supervised from time to time, and know the content of the ACOF used by their supervisor.

As we mentioned above, listening to a single call is usually insufficient for filling a form like the one reported in Table 1. In the SENSEI scenario the Quality Assurance supervisors may:

1. Benefit from automatically generated surveys for a greater number of monitored calls that, as in the manually generated ACOFs, are organized around main areas of interest, including Call Opening and Closing, Communication Skills, Problem Solving, and Proactivity.

2. Access to personalized reports obtained from the automatically generated ACOFs

3. Use those reports for focusing his/her listening task on single agents, or specific area of interest.

4. Since the ACOF items are scored, the following indicators can be reported:

    a. Overall score of the call quality

    b. Overall score of the call quality for each main area of interest

c. Overall score for each single behavior within each area of interest

d. Identification and extraction of problematic calls, i.e. the ones scored below a given threshold for 1 – 3 above.

e. Identification and extraction of calls managed by a given agent.

2.2. Conversation summaries for the speech use case

Conversation Summaries are in focus in D1.1 Use Case 3 (Reporting for the Quality Assurance Managers and Professionals – Conversation oriented summaries). The Actors of this use case are the QA Professionals and QA Managers of call centres. While studying in depth the possible uses of conversation summaries in call centres, we decided to slightly shift the focus of the summaries from agent behavior to customer behavior and call content. The rationale behind this shifting was the following:

- A set of pilot summary annotations was done by two TP QA professionals, following the Guidelines for Summary Annotation written by AMU; the annotators worked independently on ten real time conversations of their call centre. For the same set of conversations the annotators also produced ACOFs.
- We could observe that the content of the summaries could either be focused on the agent behavior and script compliance, or on the call content and customer requests.
- The first focus extensively overlapped aspects of agent behaviour already dealt with in the ACOFs, while focusing on customer and call content produced important new information like the reasons for the inbound calls, the customer reactions to outbound contacts, among others.

The Goal and the Steps of the Use Case has been reformulated as follows:

Goal: Get daily feedback in the form of reports including conversation summaries oriented to extract indicators concerning the reasons for the calls, and customer attitudes and behaviour.

Steps:

1. The QA professionals identify a call centre campaign they will need assistance with for obtaining daily reports including the reasons for the calls, customer attitudes, action required, and agent actions.

2. The system performs analytics on recorded data with respect to the specific aspects identified by the QA professionals; reports present data either in graphical form or in the form of summaries that report conversation contents related to the aspects of interest and refer to aggregations of calls based on such semantic contents.

2.2.1. Speech use case conversation summary: Scenario of use

While ACOF-based quality assurance process is in place in the call centre operations, the use of conversation summaries is completely new. For identifying the benefits deriving from them we needed a qualitative research method and we chose to run focus groups with potential users. As we wrote above, for both focus groups the goal was to collect insight about how the kind of results that SENSEI can provide could meet with user requirements emerging from the scenarios. In order to get natural conversation as well as well focused discussion, the moderator used the Interview Guide A reported in the Appendix, while Interview Guide B is the one adopted for secondo focus group. In both cases the subjects were not aware that the moderator guided the discussion on the basis the interview guides.

The second focus group was held after that summary annotation, in form of synopsis, were tried by TP front office supervisors on a set of ten conversation actually occurring in the call centre. Two annotators participated in the focus group.

The goal of the focus group was to identify possible uses of synopses in the call centre supervising activities. It turned out that 77% of the working time of agent supervisors is devoted to call monitoring and is limited by the scope and small scale the human resources devoted to the manual annotation of forms. As explained in D1.1 the listening activities are focused both on agent observation and on the need of identifying critical conversations. As a consequence call centre managers would welcome tools that can help in improving the listening activities done as part of the Best Quality Assurance process. Many call centres already rely on tools whose aim is the effective planning of the listening activities. However, those state-of-the-art tools only help in assuring that each call centre agent is monitored by a supervisor at least twice a month. It is likely that during those planned listening possible critical calls are seldom taken into account.

From the group discussion we could hypothesize that the generation of summaries for each call, and the post-hoc aggregation of calls in terms of similarities of content or occurrence of critical behavioral traits could provide the Front Office Supervisors with insight for improving effectiveness of their listening.

For example the call centre clients may require that inbound calls may be exploited as opportunities for sale proposals. However this is not always the case. The lack of sale proposals may depend on several different reasons, including the fact that the call centre agent forgot to do it, or that the call was critical because of a very unsatisfied client. From the discussion in the focus group, where this issues was repeated, it emerged that more effective listening can improve, for example, the process of identifying failures of sale proposals by focusing the listening on aggregation of calls where the sale proposals did not occur. The focus group participants considered this as a valuable contribution to the QA professionals work.

3. Social Media Use Cases

In this section we describe work we have carried out in the final phase of use case development. There were three stages: first, we revised the use cases presented in D1.1, based on further analysis and discussions with users; second, we gathered feedback on the revised use cases through more systematic consultation with users; and thirdly, we made a final prioritization of the use cases, on the basis of this user feedback.

3.1. Stage 1 – Revising the Use Cases

We reviewed the use cases, as described in D1.1, and on the basis of further analysis and discussion with user representatives (i.e. members of the public and news professionals) we made the following revisions. We simplified and improved the overall consistency of the use cases in order to communicate the ideas more clearly and effectively. In particular, we changed some terminology. For example, “actors” are now “users”; “goals” are “user goals”; the use case “steps” are now “SENSEI functionality”. We also introduced some phrases to characterize the outputs of a use case, e.g. the “Town Hall Meeting” or “THM” summary; “The Update and Extend” report; The “Comment Poster Profile”. Additionally we re-formulated the “user goals”, and these are now relevant to more user groups. Although the structure is essentially the same, some details of the use cases have been elaborated. For example, we now make it clear that users may “drill down” from summary reports to view the contributing comments.

Each Use Case is now structured into three sections: ***user groups***, a ***user goal*** and ***SENSEI functionality***.

User groups. These represent: i) various roles in the professional news media: (e.g. reporters, sub-editors, comment editors etc.), and ii) members of the public who engage with on-line news and comments. In this second group we distinguish between news and comment readers (i.e. people who read news and/or comments but very rarely post comments) and comment providers (i.e. people who read news and/or comments and who contribute reader comment on a regular basis). We note that news professionals such as reporters, sub-editors etc. may carry out activities such as reading and posting comments in their professional roles but these activities are not to be confused with the user roles of comment reader and comment provider, which we reserve for members of the public who engage with the comments. To further clarify, we view reading and posting as modes of interaction with the comments, as are skimming, searching, browsing, etc.

User Goal. This is something a user might want to achieve when engaging with news and comment.

SENSEI Functionality. A sketch of what the SENSEI software will do to help the user achieve their goal.

The different use cases describe different applications of technologies, each providing the user with a particular view of the total space of reader comments. These views may overlap. In each use case the end output includes a summary report, for specific comments and articles; and in each case we envisage links between the summary and the contributing comments, in the context of the original comment sequence.

In *Use Case 1*, the Town hall Meeting (THM) summary provides the most comprehensive view of a single set of comments and the news article. In *Use Case 6*, "Making Content from the Comments", a digest is drawn from multiple sets of comments: the output being multiple THM summaries, each drawn from a set of comments and articles -- the top ranked sets of comments from, say, a day's or a week's worth of comment and news.

In *Use Case 2*, the "Update and Extend" report, like the THM summary, provides a view of a single set of comments and article, but in *2* this is a more specialised view, based only on comments which correct or elaborate on content in the article, (it does not include expressions of sentiment or opinion on the article or a summary of the contributors).

Use Cases 3 and 4 both provide views which may cut across sets of comments by aggregating over comments which relate to a particular comment provider (as in *3*) or aggregating similar comments to a particular comment of interest (as in *4*).

Use Case 6, which produces a "Trend" report, provides a view drawn from multiple sets of comments and articles, (as in *Use Case 5*), but it displays the results over-time, and is not concerned with ranking the comments.

3.1.1. Use Case 1: The Town Hall Meeting Summary

Users:	Members of the public (news & comment reader, comment provider); News-media professional (reporter, sub-editor, comment editor).
User Goal:	To gain an understanding of the key content of a news article and associated set of reader comments.
SENSEI Functionality:	Given a news article and a set of reader comments, the system generates a ***Town Hall Meeting (THM)*** **summary** - a short, mainly text-based summary, in the style of a town hall meeting news report. A Town Hall meeting typically comprises an opening statement(s) on a particular issue(s), followed by questions and discussion from the floor. Here the news article is the opening statement and the reader comments are the public discussion. More specifically, a **THM summary** will include the following information, as if the reporter had asked the kind of questions he would use when covering a town hall meeting: • "*what's the lead here?"* ▪ a headline summary of the main story in the article and comments; • "*who took part*?" ▪ the key contributors, based on e.g. total number of posts, initiated threads, length of comments, number of direct replies, etc. ▪ the total number of comment posters; ▪ a profile of comment posters (e.g., in terms of "very frequent poster", "occasional poster", "new poster" etc.); • "*what were people talking about*?" ▪ a list of the main topics addressed in the article and comments; • "*what issues did people feel strongly about?*" ▪ the topics which were associated with intense feelings • "*what issues did people agree/disagree about*?" ▪ the topics where there was consensus / divided opinion We provide links to the comments that contribute to the different parts of the summary to allow the user to **"drill down" to see the source comment(s)** in the context of the original thread/discussion. E.g., for each of the main topics listed, there will be a link to comments associated with that topic.

3.1.2. Use Case 2: Updating/Extending the Coverage of the News Story

Users:	Members of the public (news & comment reader); News-media professional (reporter).
User Goal:	To gather from the reader comments any additional information that refers directly to the article content, e.g. reports of factual errors, elaboration, recommendations for follow up, etc., with a view to **updating the article and extending the coverage** of the news story. (Assertions of opinion, off-topic comment, etc. are excluded).
SENSEI Functionality:	The user selects a news article and a set of reader comments. The system then analyses the comments in relation to the article and gives a SENSEI "**Update and Extend**" report which includes the following information: • Claims of factual errors. • Comments which elaborate on any content in the article, e.g. by introducing new, or related facts or evidence. • Comments which propose similar examples to content in the article. • Any accounts of personal experience of issues raised in the article. • Recommendations for follow-up to issues raised in the article. Note that all of the above will include references to the relevant sentence(s) in the article.

3.1.3. Use Case 3: Backgrounding - Looking in Greater Depth at a Comment Poster

Users:	Members of the public (news and comment reader; comment provider); News-media professional (reporter).
User Goal:	To build a picture of a comment provider (aka "poster") based on other comments he/she has made.
SENSEI Functionality:	The user identifies a comment poster to the system, e.g. via one of their comments; a list of comment posters. The system returns a SENSEI "**comment poster profile**" to the reader that includes the following: • whether the poster of this comment is a prolific commenter; • the range of subject areas across which this poster comments; • whether this poster's comments typically garner many responses and whether the responses tend to be positive or negative; • whether this poster exchanges comments with many other posters, or whether s/he tends to interact with a limited set of other posters, and if so who; • a characterization of the poster and their interests in terms of the language they use, e.g. a word cloud or key phrases. The system also provides an interface that allows the reader to `drill down' to see other comments by this poster, potentially filtered by subject area, who he/she is talking to, recency, etc. and in context of the original thread. E.g., for each subject area listed, there will be a link to comments associated with that subject area.

3.1.4. Use Case 4: Finding Similar, Related or Redundant Postings

Users:	Members of the public (news and comment reader; comment provider); News-media professional (reporter; sub-editor).
User Goal:	Given a comment of interest, to find other **similar comments** (i.e. comments that make the same point or are closely related in content).
SENSEI Functionality:	The user selects a comment of interest. This may be: 1) an existing comment or, 2) a candidate comment (i.e. the user has drafted a new comment and is thinking about posting it in the on-line comments). He may choose which comments to search in - for example: • within the same article and comments; • within other articles and comments in related topic areas (e.g. "other comments on environment stories") etc. Given a comment of interest and search space; the system searches for similar or related comments and generates a SENSEI "**Similar Comment**" report. This will include: • a list of related comments ranked by their degree of similarity to the comment of interest. • simple summary statistics, such as: the number of comments and articles searched; the total number of similar comments found; total number of distinct articles and comment sets including similar comments; topic areas which include similar comments; number of distinct authors of similar comments, etc. The user can view each similar comment in the context of the original thread/discussion via links in the list of similar comments.

3.1.5. Use Case 5: Identifying Trends in Reader Comments

Users:	Members of the public (news & comment reader; comment provider); News-media professional (editor; sub-editor).
User Goal:	To determine which topic(s) in a specified time period (e.g. week, quarter, year, etc.) have elicited a significant response from the comment posting community. In particular, to identify news article topics with: very high or low volumes of reader comment; with the most emotive reader content; with the most polarized opinion and to establish what topics emerge in the comments.
SENSEI Functionality:	The user indicates a date range and selects one or more news **article topics** (e.g. "Environment", "Business", "Scottish Referendum", "Phone Hacking", "2014 Floods", etc.) For all news articles in the specified topic(s) and date range, the system analyses the associated sets of reader comments and generates a SENSEI "**Trend**" report. This will include an indication of the character of reader interest in the news **article topic(s)** based on: • the extent of comment on the articles over-time; • whether comments reflect agreement or disagreement amongst commenters; • the strength of feeling expressed in the comments and whether this is fairly consistent or highly polarised; • and finally, **reader topics** which emerge in the comments, over-time, e.g.: For a set of news articles on the topic "Scottish Referendum" the system may identify reader topics in the comments such has "NHS, devo max, freedom, Andy Murray, Westminster Elites". Then, for each **reader topic** the system indicates: • the extent of comment on the topics, over-time; • whether comments reflect agreement or disagreement amongst commenters; • the strength of feeling expressed and whether this is fairly consistent or highly polarised. We provide links to the comments which contribute to the different parts of the trend report to allow the user to **"drill down" to see the source comment(s)**.

3.1.6. Use Case 6: Making Content from the Comments

Users:	Member of the public: (news and comment reader; comment provider); News-media professional (comment editor, including letters page editor).
User Goal:	To obtain a view of the 'best' content in the comments, ranked by various criteria.
SENSEI Functionality:	The user selects a set of news articles and associated comments and a time period (e.g. a day, week, etc.). The system then generates a SENSEI "**Comment Digest**". This digest will include: (1) A SENSEI **Town Hall Meeting (THM) summary** (see **Use Case 1**) of each of **the top five** articles and comments, as indicated, e.g. by volume, strength and polarity of comment and user ratings. Each THM summary includes a headline style summary of the comments in relation to the news article. (2) A list of the top 20 comments, chosen from all comments in the time period, according to an automated ranking, as indicated by various features such as e.g.: emotiveness of comment, user ratings, number of replies, reference to new evidence relating to the article, etc.

3.2. Stage 2 – Gathering Feedback from Users

Having revised the use cases, we proceeded to gather feedback from users on the use cases via on-line questionnaires.

3.2.1. Recruiting Participants

We invited two groups of participants to take part in this study: members of the public and news media professionals. We invited members of the public via local university volunteer lists and personal contacts. We also invited news media professionals working in various roles at The Guardian Newspaper and academics in the Department of Journalism Studies, who have professional experience of working in the news media. Finally, we invited a range of alumni of the Department of Journalism Studies who are currently working as news media professionals in organisations including newspapers, radio and television broadcasting, and press and public relations for public services.

3.2.2. On-line Questionnaires

We designed two main questionnaires: the first to collect background information from participants and the other to collect feedback from the participants on the use cases. For each of these questionnaires we produced two variants, one for members of the public and the other for news media professionals (see Appendix D).

We published the questionnaires on-line, which allowed us to gather responses remotely and anonymously. We presented the questionnaires together with an introduction to the project, some simple instructions on the questionnaires and the participant information forms. (The latter are required by the University of Sheffield Ethics Procedure – see D8.3). A demo of the complete on-line questionnaire for news professionals may be viewed via: http://paramita.staff.shef.ac.uk/sensei/demo/introduction.php

Details of the forms follow:

Background Questionnaires. These invited participants to provide information on their past experience of using on-line news and reader comment. As stated above, we produced a version for each of the user groups: i) for news media professionals and ii) for members of the public. We designed questions with a view to exploring patterns of use and experience within the different groups and roles. We included a question asking people to identify their 'role' within the respective groups – news professionals (e.g. reporter, editor, comment editor etc.) and members of the public (e.g. news and comment reader, news and comment provider). A key difference between the questionnaires is that in i) we ask people to provide responses mainly on the basis of their experience of using news and reader comment in a professional context (i.e. as part of their work). Whereas in ii), we invite responses based on personal use of news and reader comment.

In both background questionnaires we included simple multi-choice questions to collect quantifiable responses about user behavior and past experience, e.g., we asked when people first started using on-line news and comment, how many news stories they typically read, how often they engage with comments, how many comments they read and provide and what are their patterns of posting, e.g. do they ever post? do they tend to post in response to something in the article? do they typically initiate a thread? In addition, we asked people to tell us a bit about why they read and post reader comments.

By linking the background questionnaires with the main use case questionnaire, we allow for an enquiry into whether different user groups show preferences for different use cases.

Use Case Questionnaires. In the main questionnaire we presented each use case in turn together with 4 questions, each designed to capture the participant's feedback and thoughts on different aspects of the use case. We used a mix of quantifiable questions and open comment style questions.

Two questions focused on the specified user goal and functionality—is the user goal realistic and is the functionality something that the users would find useful? In these we formulated two respective statements, and used a 5 point Likert scale to capture how much people agreed with the statements, in respect of the different roles within the user groups.

We also wanted to gather evidence of the task contexts in users might employ the proposed functionality. Such information is helpful in prioritizing use cases, and may also inform work on extrinsic evaluation for social media technologies in SENSEI. Providing a few indicative examples, we invited participants to tell us about any activities in which they would use the functionality and to tell us why they thought it would be useful.

In a further question, we invited people to tell us anything they would like to modify or add or remove from each use case.

After reviewing all the use cases, a final question asked participants to order the use cases from most preferred to least preferred, and to give any further general comment.

3.2.3. Results from On-line Questionnaires

The results and initial analysis presented here are based on responses from 31 participants, including 18 members of the public (16 news and comment readers and 2 comment providers) and 13 news-media professionals. Of these, 18 completed the full questionnaire (11 members of the public and 7 news media professionals). As Table 6 in Appendix E shows, the number of participants decreases as they progressed through the questionnaire, for example 31 participants had completed up to and including the questions for Use Case 1 (18 members of the public and 13 news professionals), but only 18 completed the entire questionnaire.

Of the news media professionals who took part, current roles included 3 reporters, 1 sub-editor, 3 executive/editorial, 2 educator/academic, 1 public relations, 1 private sector worker, 1 communications officer and 1 social media editor (See Table 7, Appendix E for a full breakdown of roles). However, we also asked participants about previous roles they have held in the media. Based on answers to this question we were able to determine that our participants included 13 who have worked (or currently work) as a reporter, 6 as a sub-editor, 1 as a comment editor, 4 as comment moderator, 8 as executive/editorial and 2 as academics.

Results are presented in Figure 1 - Figure 6 in Appendix E. Figure 1 and Figure 2 show the responses to questions 1 and 2 respectively across the six uses in the Use Case Questionnaire[1], distinguishing responses for the two broad user categories of the public and the news professionals. Figure 3 - Figure 6 all relate to the final question where participants were asked to rank the use cases based on their perceived usefulness. Figure 3 shows for each use case the number of respondents in the news professional category who placed that use case in each of the 6 possible rank positions. Figure 5 does the same thing for responses from members of the public. Since that level of detail is hard to interpret, we re-present the same data in Figure 4 and Figure 6, where instead of 6 ranks we simply have two ranks: “high ranks”, comprising ranks 1-3 in the initial ranking and “low ranks”, comprising ranks 4-6 in the initial ranking.

3.2.3.1. Preliminary quantitative analysis

Overall results from participants (professional and members of the public) show a favourable response to all six use cases. The average scores for the questions investigating 1) the perceived authenticity of the “user goal” and 2) the perceived usefulness of the functionality were higher than the midpoint of 3 on the Likert scale for all 6 use cases (see Figure 1 and Figure 2 in Appendix E). We note that the difference between the lowest and highest average scores for the individual use case functionalities rated in question 2 was less than 1 point on the Likert scale. Moreover that as Table 6 shows, more participants answered questions for the first 3 use cases (31, 27, and 22 respectively), and with people failing to complete the entire questionnaire, only 19 provided answers for Use case 6 and 18 completed the final ranking

[1] For each use case, participants were asked 1) if they agree that the user goal (for the particular use case) feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments (*1=strongly disagree, 5=agree*), and in 2) if they think that SENSEI functionality would be very useful for different user groups, when they engage with on-line news and comments (*1=strongly disagree, 5=agree*).

question. Thus comparing average scores on Use Case 1 with those on Use Case 6 should be done with some degree of caution.

The final question, where participants were asked to rank the use cases based on their perceived usefulness, does not indicate any absolute notion of usefulness – just usefulness relative to the others – and hence should not be interpreted as indicating that the lowest ranked use case is not perceived to be of value.

The Public – what people liked

Overall, members of the public who completed the questionnaire allocated the best overall ranking position to the functionality in Use Case 6 – the Comment Digest report. More than half of responses (6 of the 11) ranked this first of six, where rank 1 = "would find the functionality most useful". This ranking is consistent with the average scores for question 2, which asked people to assess the utility of the proposed functionality2. As Figure 2 shows, the highest average score (just under 4 on the 5-point Likert scale) is for Use Case 6, the Comment Digest functionality.

The public showed very little difference in their responses to and overall rating of the functionality proposed in Use Cases 1, 2 and 5 (Figure 2). In terms of the ranking, all of Use Cases 1, 2 and 5 had on balance more respondents placing them in high ranks than in low ranks, with slightly more preferring Use Cases 1 and 2 to Use Case 5 (Figure 6). In fact Use Cases 1 and 2 were placed more in high ranks even than Use Case 6, though it was ranked first by most participants. This suggests a broad consensus that Use Cases 1 and 2 are useful. By contrast, Figure 6 clearly shows a lower preference for Use Case 3 and 4 (the Comment Provider Profile and the Similar Comments Report).

News Professionals – what people liked

While overall members of the public liked Use Case 6 best, the news professionals ranked Use Case 5 most highly -- four of seven respondents placed it first and all seven ranked it in the top three (Figure 3 and Figure 4). Following this, the next two most highly ranked use cases were Use Case 1 and 2. The other three were all placed in low ranks by more participants than placed them in high ranks (Figure 4).

3.2.3.2. *Preliminary qualitative analysis*

Participant responses to the comment style questions[3] were mostly positive. There were even a number of explicit commendations for the proposed functionality and ideas, which included comments such as "good idea… would be very useful", "the idea …would be great"; "excellent idea "; "pretty well done".

[2] See e.g. question 2 for Use Case 6, which asked "Please indicate how much you agree with the following statements - A SENSEI "**Comment Digest**" would be very useful for the user groups, when they engage with on-line news and comments".

[3] Based on the "comment style" responses to questions 3 and 4 (q.3. asking people to say in what context and why they would find the functionality useful; and q.4. asking people to give details of what they liked disliked etc.) and not including examples that we provided.

What participants said they particularly liked/would find useful in respect of the different use case functionalities and their potential (where numbers in brackets indicate the use cases for which the point was raised)[4]:

- An overall summary of what people are talking about/the main arguments in the comments (1, 5 and 6)
- Discovery of "hot topics" (1, 5 and 6)
- Taking the emotional temperature of the debate/ topics in the comments (1, 5 and 6)
- Functionality that identifies the more factual comments, relevant to the article (2)
- Fact checking/suggestions for corrections to the article (2)
- Having reliable recommendations for comment/content (6)
- Having help to find ideas for follow on stories (1, 4, 6)
- Having providers be more visible/accountable for their comments (3)
- Assembling background/context for assessing sources (a comment/the author of a comment) (3, 4)
- Finding candidate comments for including in a piece or for writing opinion pieces (3)
- Finding/filtering spam (2, 3)
- Checking if someone has posted a similar post to yourself (4)
- Finding comments linked to sources of evidence (2, 5)
- Having commenters "typed" in some way e.g. "demographic" (1, 5)
- Time saving functionality (1, 4, 6)

By contrast to the above, the more notable concerns raised in the participant responses for different use cases include:

- Having insufficient time to engage with the proposed functionality (3, 4)
- The undermining of the anonymity of providers and its implications for the character of the debate (3)
- The potential for certain comment providers to exploit the proposed functionality with a view to deliberately distorting i) the overall view of the comments and/or ii) what was said in the article (1, 2).
- Focusing on measures of "emotive" comments-- the functionality may risk presenting a polarised or sensationalised view of the debate. (5, 6)
- How far one could vet/trust the authority of what was said in the comments. (1, 2)

[4] Again, these findings were based on participants' own free comments.

3.3. Stage 3 – Recommendations for Use Case Prioritisation

Informed by a number of considerations we have decided in SENSEI to first pursue use cases 1 (Town Hall Meeting Summary), 5 (Trend Report) and 6 (Comment Digest). Other use cases, particularly use case 2 (Update and Extend Report), will be considered if there is time. Factors influencing our decision are:

1. The quantitative results from questions 1 and 2 on the questionnaire. As discussed above these indicated that the public most highly valued use cases 1, 2, 5 and 6 while the news professionals most highly ranked use case 1, 2 and 5.
2. Free text comments supplied by questionnaire participants, which revealed, in contrast to the averages derived from the quantitative results, where participants were particularly excited by functionality, or the potential for the functionality offered in the use cases. In particular, see the first three points about what participants liked in the use cases made above in the qualitative analysis section. Also while there were some concerns about all the use cases, a considerable number of participants shared the concern that engaging with the functionality proposed in Use Cases 3 and 4 would simply take more time than users would be prepared to spend.
3. Technical considerations about difficulty of implementation and possibilities for reuse/sharing of underlying capabilities across multiple use cases. Use cases 5 and 6 build on use case 1 and can reuse substantial amounts of the functionality that will be required to implement use case 1. By contrast, use case 2 (Update and Extend Report), while popular will require the development of additional challenging capabilities to determine (a) which comments relate to claims in the article and what claims these are and (b) whether a comment is updating or extending a claim in the news article.

Taking these factors into account, we believe Use Case 1 is where SENSEI should start with development and that following an initial implementation of use case 1 effort should be split between gathering and ranking Town Hall Meeting Summaries for multiple articles (Use Case 6) and clustering top comments from multiple articles over time to determine the emergence of trends (Use Case 5).

Finally, while our respondents made many helpful comments regarding aspects of the use cases that should be borne in mind as we move forward towards implementing them, there were no suggestions for major revisions to the use cases. Therefore the descriptions of the use cases presented above in Section 3.1 stand as the specification we intend to take forward in the next stages of the project.

4. The SENSEI Evaluation Methodology

In this Section we pave the way for D1.3 (the first SENSEI evaluation report), and we describe the evaluation methodology and metrics we will use. The evaluation model includes technology-oriented (intrinsic) metrics, task-oriented (extrinsic) metrics and insight-oriented metrics. Insight-oriented and Extrinsic metrics are qualitative measure, while the nature of intrinsic metrics is quantitative. That difference is reflected on the evaluation baselines that we illustrate in the last section of this chapter.

This section is organized as follows. The first paragraph (4.1) includes the descriptions of metrics that are applied at the technology level to assess how SENSEI technologies developed in WP4 and WP5 perform. The second paragraph (4.2) illustrates intrinsic evaluation scenarios for the SENSEI prototype, while extrinsic evaluation scenarios are described in paragraph 4.3. Both intrinsic and extrinsic evaluation scenarios are specified with respects to the speech and social media domains. Finally paragraph 4.4 describes the insight-oriented analysis that are envisaged for the new activities enabled by SENSEI analytics technologies

4.1. Technology level metrics

4.1.1. Metrics for the Semantic Parsing evaluation (WP3)

In this section, we introduce metrics for evaluating the effectiveness of automatic frame annotation of conversational data. This task goes through several steps, which include A) Frame Selection, B) Argument Boundary Detection, C) Argument Semantic Labeling, D) Frame-to-Frame Relation Annotation. It is relevant to evaluate the above steps both individually and as a whole. We define a set of seven incremental evaluation metrics which is hereby detailed. Each metric is expressed in terms of **Precision, Recall**, and **F-measure**. Here, the allowed boundary of a frame’s textual realization (including all of its arguments) is assumed to be an individual dialog “turn”, either a speech turn or a social-media message. Not all metrics are going to be used for every use-case in the intrinsic evaluation processes of WP3. Depending on the availability of reference annotations, we will use only some of these metrics, although a full evaluation on a limited subset of each dataset will be provided at the end of the project.

4.1.1.1. Preliminary Definitions

Frame: abstract representation of an action, event, or property appearing in a sentence (e.g. "Awareness", "Expensiveness")

Argument: a conceptual entity involved as participant in a Frame (e.g. "Goods" as a participant in the "Expensiveness" frame)

Example: the text segment "good morning I would like to know how much a double room costs" includes realizations for both the Awareness and the Expensiveness frames. The former is related to the verbal predicate "know", while the latter is related to "costs".

4.1.1.2. Metrics

1. ***Frame Recognition***. Evaluates the performance in recognizing the presence of individual correct frames (disregarding their related Arguments). Example: recognize that the frame "Expensiveness" is present in the sentence.

2. ***Argument Boundary Detection***. Evaluates the performance in detecting the exact word-level boundaries for individual Arguments, disregarding their labels. Example: detecting that "a double room" is an argument. Actual word boundary detection technique may depend on the quality of transcription (automatic/manual). When performing measurement on automatic transcriptions, it might be necessary to allow a controlled boundary approximation, or equivalently to introduce a process of word-level normalization when mapping actual words to Arguments.

3. ***Argument Labeling*** (Frame-to-Argument links). Evaluates the performance in detecting the correct argument labels, disregarding their boundaries. This means correctly enumerating just the list of arguments which are present for each frame in the sentence. This is equivalent in detecting links between Frames and their respective Arguments. Example: detecting that a (not specified) Argument of kind "Goods" is present.

4. ***Argument Recognition***: evaluates the composition of the tasks defined in the above metrics 2 and 3 (shortly “2+3”), assuming the task defined in metric 1 as given. Therefore, Argument Recognition valuates the overall capability of annotating Arguments, in isolation from Frame Recognition. Example: we know the "Expensiveness" frame, and detect that "a double room in Ibis" is a "Goods"-kind Argument.

5. ***Frame Realization***: composition of 1+2+3, so the metric evaluates the capability of annotating correct individual frame instances, including Frame Labels (1), Argument Boundaries (2), and Argument Semantic Labels (3).

6. ***Frame Composition*** (Frame-to-Frame links/relations). Evaluates the performance in detecting the correct Frame relations, e.g. the Frame Awareness is linked to the Frame Expensiveness. This is typically and "inclusion" relation, e.g. the Expensiveness Frame acts as a whole Argument of the Awareness Frame.

7. ***Turn Analysis:*** composition of 5+6, so each conversational turn annotation is considered correct if and only if all of the compounding steps are completely correct.

4.1.2. Metrics for the evaluation of intra-document coreference

4.1.2.1. The task and its terminology

Intra-document coreference resolution is the task of identifying which segments of text (e.g., noun phrases, or NPs) are mentions of the same (discourse) entity.

In the example text in (1), for instance, NPs John, He and he are mentions of the same entity, whereas Mary and she are mentions of a second entity, and Canada of yet a third one.

(1) John1 met Mary2. He1 told her2 he1 was moving to Canada3.

Equivalently, one can say that in this example we have three equivalence sets of mentions or **coreference chains**: the three mentions of entity 1, the two mentions of entity 2, and the solitary mention of entity 3.

In the case of intra-document coreference , therefore, the gold standard G is therefore a set of entities whose associated coreference chains G1 .. Gn specify a partition over the set of mentions:

G = G1 .. Gn

4.1.2.2. Metrics for evaluating Intra-Document Coreference

A number of metrics have been proposed to evaluate Intra-Document Coreference. We discuss each in turn.

The MUC metric. The first widely used metric for evaluating coreference was proposed by Vilain et al. (1995) for the 1995 of the Message Understanding Conference (MUC) evaluation campaign, and became known as the MUC scorer.

The MUC scorer is link-based, in that it computes precision and recall for the output of a system (the response) by looking at each coreference chain in the gold standard G = G1.. Gn or the response R = R1.. Rm as a connected graph. A completely correct response will consist of all and only the links required to connect all the mentions in the equivalence sets of mentions G. A correct link is a link in the response R between two mentions that are part of the same coreference chain in the gold standard. Recall is then defined as the ratio of the number of correct links in a system's response over the total number of links in the gold standard. This total number of links in the denominator is the sum of the number of links required to connect each of G1..Gn:

$$\sum_{i=1}^{n} |G_i - 1|$$

And it can be shown that the total number of correct links regarding gold standard coreference chain Gi in the numerator is equal to the size of Gi minus the partition of Gi induced by the response R, p(Gi, R)–the number of sets in which Gi has been split in the response because the system missed a coreference between mentions:

$$\sum_{i=1}^{n} |G_i - p(G_i, R)|$$

Thus recall in the MUC scorer is defined as follows:

$$R_{MUC} = \frac{\sum_{i=1}^{n} |G_i - p(G_i, R)|}{\sum_{i=1}^{n} |G_i - 1|}$$

Precision in the MUC scorer is defined by reversing the roles of the gold standard coreference chains Gi and the coreference chains in the response Rj . Clearly, a link in the response is incorrect if it connects two coreference chains that in the gold standard are not connected. Thus, the number of correct links in the response can be calculated by subtracting the 'extra' incorrect links linking chains that shouldn't be–whose number is given by the size of the partition induced on coreference chain Ri by gold standard G.

$$\sum_{j=1}^{m} |R_j - p(R_j, G)|$$

Precision in the MUC scorer is then defined as follows:

$$P_{MUC} = \frac{\sum_{j=1}^{m} |R_j - p(R_j, G)|}{\sum_{j=1}^{m} |R_j - 1|}$$

B3. The B3 metric, proposed by Bagga and Baldwin (1998), attempts to overcome the limitations of the MUC scorer by focusing on mentions and computing mention-based precision and recall measures. Suppose mention mk belongs to coreference chain G1 in the gold standard and to coreference chain R1 in the response. Then recall for mk is

$$R_{B^3}^{m_k} = \frac{|G_i \cap R_j|}{|G_j|}$$

whereas precision is

$$P_{B^3}^{m_k} = \frac{|G_i \cap R_j|}{|R_j|}$$

Overall precision and recall are the average of the precision and recall scores for the individual mentions.

B3 doesn't suffer from the problem of singletons but still excessively rewards systems that connect all mentions in a single coreference chain (they get 100% recall) or that assign each mention to a separate coreference chain (they get 100% precision).

CEAF. When computing precision and recall in B3, a coreference chain R1 in the response can be associated with different coreference chains in the gold standard Gi, Gj, etc if the mentions in Rjbelong to different chains in the gold standard. The CEAF metric, proposed in (Luo 2005), was designed to prevent this 'multiple match': the computation starts by specifying an aligment such that each coreference chain in R is associated with one and only one coreference chain in G. In this sense, CEAF can be said to be entity-based–it is based on an alignment between the entities. Given an optimal alignment g between G and R (i.e., one which maximizes overlap) precision and recall are then computed by measuring the similarity φ between response entity Rj and the entity g(Rj) in the gold standard associated to Rj by g.

Different versions of CEAF are possible given different definitions for φ: e.g., by defining φ(R, S) = R ∩ S we obtain what is called CEAF□3 .

Evaluation campaigns and MELA. All the metrics discussed above capture plausible intuitions about evaluating coreference. But the results do not tend to converge, as illustrated most clearly by SEMEVAL Task 1 on Multilingual Coreference (Recasens et al. 2010), where almost every system came on top according to a particular metric. In subsequent evaluation campaigns, therefore, the MELA metric proposed by Denis and Baldridge (2009) was used, which takes a weighted average of MUC, B3, and CEAF. In particular this metric was used for both the CONLL-2011 and CONLL-2012 shared tasks (Pradhan et al 2012, 2013).

We propose to use as evaluation metric for intra-document coreference in SENSEI MELA, which is also implemented as part of BART's evaluation package.

4.2. Intrinsic Evaluation Tasks

4.2.1. Speech Use Case Intrinsic Evaluation Tasks

We plan on evaluating SENSEI technology that can (1) generate synopses of call centre conversations, (2) automatically fill ACOF questions. The following describes evaluation metrics for those two tasks.

Synopsis generated by the system will be compared to gold standard synopses written by human experts (QA professionals). Following work by the summarization community, we can devise two types of evaluation: manual rating of the form and content quality of the synopses, and automatic comparison of synopses with gold standards. For the first type of evaluation, human judges should read the system output and rate them on a 1-5 scale, according to the following questions:

- Readability/fluency: is the synopsis written in readable language? This aspect might also be evaluated automatically with metrics like GLEU (Mutton et al, 2007).

- Content: does the synopsis cover relevant material of the conversation? This aspect might also be evaluated with the Pyramid method (Nenkova et al, 2004) which consists in listing relevant aspects in reference synopses, and manually matching them in the system synopses.

The second type of metric consists in comparing system synopses with multiple human-written gold standard synopses. This evaluation will be performed with ROUGE, which is the standard in the summarization community, keeping in mind its limitations. ROUGE (and variants) is the number of items (generally word n-grams) that overlap between the system synopsis and a set of gold-standard synopses, divided by the number of items in the set of gold-standard synopses. ROUGE also often includes the computation of average results over randomly sampled subsets of reference synopses, in order to get better estimates of the metric. We plan on contrasting ROUGE over word n-grams with ROUGE over semantic frames as detected by WP3 tools, similarly what is achieved with ROUGE-Basic-Elements with syntactic dependencies.

Evaluation of textual productions such as summaries, is a very difficult task, because it is impossible fully describe a gold standard summary, which entails that two experts will inevitably write different summaries because they may focus on different aspects, or use different wording. For this reason, the set of metrics proposed here have been shown to have limitations. Manual evaluation on a Likert scale is expensive, opaque and subject to expert disagreement; The pyramid metric relies on expert-annotated SCU (summary content units) which are very difficult to define at a correct granularity, and the matching between SCUs and system summaries by experts is subject to disagreement, or if performed automatically, subject to entailment errors. The family of ROUGE metrics has also been shown to be very limited (Sjöbergh 2007): systems can cheat to output high-scoring ROUGE summaries which don't make sense to a human (for instance by using a bigram word model trained on the source documents and outputting the highest probability sequence of words respecting the length constraint). Even though high when aggregating all runs of a system, the correlation between ROUGE and manual evaluation collapses when computed at the topic level. For those reasons, we plan on not trusting blindly the metrics we will use but rather study their behavior and propose alternatives where possible.

The ACOF filling task consists in automatically generating yes/no/na labels for each of the ACOF questions, as defined in 2.1.3. This task is a classification task and can be evaluated with the accuracy metric (number of correct answers over number of trials) as well as the precision, recall, f-score triplet (precision is the number of correct hypotheses over the number of hypotheses, recall is the number of correct hypotheses of the number of references, and f-score is the harmonic mean between the two). There are additional specificities with the ACOF filling task. It is not established that experts creating a gold standard for this task shall get a perfect inter-annotator agreement; on the contrary, one expects variability in judgments which have to be accounted for in the evaluation. For this reason, we will compute inter-annotator agreement on a subset of the corpus annotated by multiple experts and we will develop specific metrics for accounting for this source of variability and contrast them to the classical performance measures. The second specificity comes from the fact that that ACOF can be filled at the agent level or at the conversation level. We will compare aggregate agent-level ACOFs with mean conversation-level ACOFs from a sample of conversations of a given agent.

For both the synopsis generation task and the ACOF filling task, k-fold training will be applied over the corpus (with at least 10 folds) in order get more accurate results. For synopses, between two and four gold-standard synopses are being collected (see D2.2 for details).

4.2.2. Social Media Intrinsic Evaluation Tasks

The objective for intrinsic evaluation is to assess how well a system or component technology is able to carry out the task for which it was designed. I.e. how close to its target output is a system able to get?

For the social media summarization subsystem (WP5) in SENSEI we are planning to develop a number of component technologies. Each of these components needs to be evaluated, as well as the summarization subsystem itself. These components are in addition to components developed in WP2, WP3 and WP4 whose outputs may be used in the summarization system and whose evaluation is discussed above.

Comment Linking. Reader comments generally relate either to a claim in the news article or to a previous comment. Establishing which prior sentence or sentences a comment relates to is a key part of topically and rhetorically structuring the comments, itself a key part of summarization. The social media summarization system will contain a comment linking component and provision needs to be made to evaluate it. The task of comment linking is also likely to form the core of the shared task (WP7) and the evaluation setup and evaluation data will be shared between the shared task and this WP5 component, if appropriate.

Gold standard data will be created by human annotation. This may be done via crowd sourcing and according to the annotation guidelines being prepared for the shared task. The baseline will be linking all sentences in non-thread initial comments to the comment the author is replying to and sentences in thread-initial comments to the article sentence with which they have the most lexical overlap or vector space similarity. The evaluation measure will be proportion of comments correctly linked.

Topic Clustering. Another core component of the social media summarization system will be a topic clustering component which will cluster reader comments and article segments that are topically related. To assess topical clustering we will assemble sets of human topically clustered article segments plus conversational threads. Various baselines will be considered. One will be to assume that every paragraph in the news article is a cluster and then link comments to these clusters, using lexical overlap or vector space similarity, as described for comment linking above. Another will be to assume i) each paragraph in the article and ii) each thread is a separate cluster. For assessing clustering quality we will use standard measures, such as purity, mutual information, rand-index and f-measure [Manning et al., 2008].

Summarization. Summaries are difficult to evaluate because no single gold standard exists for system output. To evaluate summarization system output, multiple reference summaries are typically produced by humans and then one of two approaches is followed. Either (a) an automatic evaluation metric, such as ROUGE [Lin, 2004], is applied, which compares word n-grams from system outputs with a set of human-written reference summaries, or (b) humans rate summaries by counting summary content units that occur in the reference summaries and differentially weighting them based on how many reference summaries they occur in (e.g. the Pyramid method [Nenkova and Passonneau, 2004). Readability evaluations may also be used to assess linguistic quality of the summaries. ROUGE-like metrics have proved to be limited predictors of human judgments and easy to trick [Owczarzak and Dang, 2011], but are easy to use repeatedly. Pyramid-like evaluations are more reliable, but more expensive to carry out.

For social media summarization we will carry out two sorts of evaluation, the first involving a comparison of system summaries with reference summaries; the second involving an assessment of system summaries based on various readability criteria. We describe these different approaches to summary evaluation in turn.

First we will create multiple reference THM summaries to be used as a gold standard against which the performance of the summarizer will be judged using the ROUGE metric. An initial, pilot collection of these has been gathered as MS1. It consists of human authored summaries for the chronologically first *k* threads of 20 news articles, such that at least 100 comments were collected per article (the number of threads needed to meet this condition varies per article). We developed a method for writing summaries and tools to support annotators using this method. Both method and tools were used in the creation of the pilot collection. D2.2, section 3.10, provides more details of this pilot collection, including: the data used, the annotations collected and the tools that support the annotation process. Here we say a bit more about the method that we used in creating the pilot model summaries. Examples may be found in Appendix F.

To date writing summaries for reader comments has been identified as a very difficult task for humans due to the character and structure of social media reader comments. We have developed a **guided human annotation task for helping humans write summaries of reader comments** in social media, and in particular, on-line news. We believe this to be an important and novel contribution to the field.

In our task, the emphasis is not on a strict definition of what the summary should look like—we do not provide a model summary template. The guidance is in the form of a method on *how* to produce a summary and is to help annotators engage systematically with the data and to help them ground their summary in the comment data, resulting in representative summaries, which are similar for different annotators.

Given a news article and a set of reader comments**,** the social media summary writing task involves a number of key stages that the annotator is encouraged to engage in as he/she develops the core content for a summary of the comments. Note these stages are not carried out in strict succession, but rather as and when the annotator feels appropriate, and they may be repeated throughout the task.

1. **Comment labeling –** the annotator proceeds through the comments labeling the comments in turn. A label is a 'short descriptive note' or abbreviation, which captures the ‘gist’ of the comment and helps an annotator to engage with the comment so that he/she recognises similar content in other comments. Labels may be refined throughout the task.

2. **Comparing comments –** as the annotator proceeds through the comments, he/she should compare content in the comment and its label with content in previous comments and labels. For similar content, we encourage the use of similar labels.

3. **Grouping labels and comments** – the annotator may choose to gather together groups of similarly labelled comments and to divide groups of similarly labelled comments into sub-groups (either mentally or using a text editor). Grouping and sub-grouping of comments and labels may go on throughout the task. This stage goes hand-in-hand with the **reformulating of labels**.

4. **Re-formulating labels** – as the annotator proceeds through the comment set, he/she develops the content indicated by the labels into more fully formulated "topics or propositions". Many of these will go into the end summary.

5. **Quantifying –** the annotator may count the number of comment providers who support the ideas represented by a particular label. Again this process is on-going throughout the task.

6. **Selecting/saving interesting comments-** the annotator may select comments which are particularly striking, on their own or in the context of other comments.

7. **Writing the summary** – the annotator assembles together the more formulated labels and quantifiers and produces a written summary, aiming for a length of 150-250 words. Again the annotator may engage in this process throughout the task. We provide an example summary to illustrate the kind of summary we expect from this data-oriented method for writing summaries.

The second sort of social media summary evaluation will be a readability evaluation carried out over a random selection of system-produced summaries. Criteria assessed will be those defined for the readability assessment in TAC. The baseline summarization system will be one that takes the first n sentences of the article and the first comment of the (chronologically) first m threads up to a predefined summary length (we presume separate length limits of the article summary and comment summary).

4.3. Extrinsic Evaluation

In this section, we describe an ecologically valid evaluation of the proposed advances in conversation understanding pursued by the SENSEI project. The ecological validity of an experiment requires approximating as closely as possible real world processes and methodologies, as opposed to lab conditions. In SENSEI, we strive to validate the usefulness of the approaches we are proposing by evaluating them in real-world scenarios drawn both from the speech use case and the social media use case. This means that for the first domain, we need to measure the impact of conversation analysis technology in call centres, and more particularly in the daily work of QA professionals. In the second domain, social media, we will evaluate that impact on the work of journalists and/or members of public who read or comment on on-line news.

In the framework of a technical advancement, it is not easy to obtain ecological validity because the proposed technology is often not replacing an existing process but rather enabling novel approaches by raising technical barriers (this issue was outlined in D1.1, especially for the social media use case). For instance, we expect SENSEI to allow call centre QA professionals to analyse a much larger number of conversations than they currently do. Similarly, we hope to empower journalists and news readers with finer and higher coverage analytics tools than they have access to. Therefore, in our extrinsic evaluations, we will make our possible to both address existing processes, and introduce new ways to solve existing tasks, for both use cases.

The envisioned evaluation consists in having subjects perform a task with and without SENSEI technology, and measure their success rate in both conditions. It is different from intrinsic evaluation which compares the output of a system to a gold standard without measuring how this system affects human behaviour. As discussed earlier, the project will also run intrinsic evaluations on meaningful tasks. Evaluation scenarios are selected according to the following criteria:

- The task is relevant to the users targeted by the use-cases and represents an added value to them.
- The task can be easily performed by ahuman user, but is non-trivial for computers.
- The task can be performed with existing technology or with SENSEI-provided technology.
- The performance of the subjects can be evaluated.

The outcome of the evaluation is a statistical measurement of whether subjects perform the task better with or without SENSEI. And since human trials are expensive, careful design must be applied in order to limit the sources of variability which can affect the statistical sample and which might confound the interpretation of the results. The sources of variability that have to be accounted for are:

- Subject ability to perform the task: some subjects have more training, experience, or relevant background knowledge or have a better understanding of the technology or learn to use it faster. Training subjects before trials can help limit this variability but cannot eliminate it. Using a latin square design can help with this issue.
- Evolution of subject ability: if subjects are used in multiple trials, they have more experience with the task and the system each time. Again, using a latin square design can help with this issue.
- Trial sample: in order to improve the generalizability of the results, a task might be run in various conditions, which might be more or less difficult for the subjects. Balancing the difficulty of trials, and calibrating them can help with this issue.
- Environment: subjects are affected by their mood, the room in which they perform and the hardware they use. Sessions should be run in short timeframe and at similar location using identical fixtures.

In SENSEI, ecological evaluation will be supported by the development of a prototype which is described in D6.1. In order to facilitate the design of this component, we list a few evaluation scenarios from which we aim to draw when running the evaluation. For each scenario, we give a description of the task, how it will be evaluated, its relevance and an assessment of its ecological validity.

4.3.1. Speech Use Case Extrinsic Evaluation Tasks

For the speech use case, we will pursue at least one of the following scenarios.

1. Agent observation form

Subjects have to analyse a conversation and fill an agent observation form covering aspects related to a call, the caller and the agent behavior, which could be: call opening, problem identification, problem fixing, increased value, summary confirmation and call closing, communication skills.

This evaluation scenario is the most ecological because the task is already performed by QA professionals in call centres.

Evaluation setup: subjects perform their task given a set of conversations involving a given agent. They have a time limit which does not allow them to just listen to all the calls. For each question of the form, they have to select one of the possible answers (yes, no, not assigned) and justify that answer, for instance by writing in their own words why they selected it, or by pointing to material in the source data which supports it.

We create two conditions:

- Baseline: the control group has access to existing technology for performing the task, such as conversation transcripts and a text search function.
- SENSEI: the target group has access to synopses, machine-filled ACOF, advanced search, powered by SENSEI-proposed approaches.

Assessment: Subjects are rated according to how well they fill the questionnaire in the given time, according to a gold standard created without a timing constraint.

2. Collection-wide conversation retrieval

Subjects perform a collection-level information retrieval task. For example, the task consists in finding conversations where the caller objectives were not met, or where the agent gave cynical replies, or where the caller was having negative emotions.

This evaluation scenario is ecological because it matches the need to analyse a whole corpus instead of a small subset as is the case in most call centres today.

Evaluation setup: subjects have to find conversations that respect a non-trivial criterion, and can use a subset of the amenities offered by SENSEI. Subjects have to explain in their own words why a given item is relevant and are rated on the number of relevant items they can find under a time constraint.

We create two conditions:

- Baseline: Subjects have to perform the task with existing technology, such as conversation transcripts and a text search function.

- SENSEI: Subjects can use synopses, semantic, para-semantic and structural analyses, as well as enhanced conversation views, and advanced search, powered by the SENSEI-proposed approaches.

Assessment: subjects are evaluated according to how many relevant conversations they can find compared to a gold standard created by experts without a time limit.

4.3.2. Social Media Candidate Extrinsic Evaluation Tasks

We will carry out at least one of the following extrinsic evaluation tasks. Others will be considered if time and resources are available.

1. *A comment editor preparing a summary of the contents of a news article and associated comments.*

This task is currently carried out on a very small scale (several articles per week) by Guardian staff. This means the task is “ecological” in the sense that it is currently being done and hence serves as a good choice as an extrinsic evaluation task for SENSEI.

Evaluation setup: participants are asked to carry out a task, i.e. write a summary of a news article and a set of comments within some overall time limit (may finish earlier if they wish).

We create two conditions:

- Baseline: current practice, where they have access to the article and comments in a typical threaded comments interface.
- SENSEI: where they have access to a SENSEI THM summary (Use Case 1) in addition to the article and comments being displayed in a typical threaded comments interface.

Assessment: Assessment will be carried out using a variety of measures and approaches that will include some or all of:

- time taken to complete the task (ideally maximum time allowed would be varied under the two conditions and the effect on the following measures observed);

- an assessment by a third party of the *quality* of the summaries produced under the two conditions according to either (1) a set of criteria, measured on a Likert scale, such as *accuracy* (does what is said in the summary accurately reflection what is said in the article/comments?), *breadth/coverage* (does summary include all the main topics of the article/comments?), *interestingness* (how interesting/engaging is the summary?) or (2) by ranking the summaries relative to each other.

- possibly, assuming feasibility, we will investigate traces left by the summary authors of the process they followed in preparing the summary, e.g. either (a) session logs of participant interaction with the news article, comments and, when available, the THM summary or (b) user-generated logs of materials viewed/gathered during the course of summary creation. Here the aim is to explore and ideally quantify differences in user interaction with the source materials in the two conditions in order to understand how SENSEI tools may differently inform the character of the output.

- a post-hoc questionnaire – following completion of the task, participants will be asked to answer a post-hoc questionnaire about their experience. This will ask them to give their subjective perception of how factors such as depth (how deeply they engaged with the comments) and coverage (how broadly they engaged with the comments) differed across the two conditions, whether the topic-based structuring of the THM summary helped in writing their summary, etc. (Note: While participants will be asked to write summaries under both conditions, they will not be asked to write summaries of the same article plus comments under the two conditions – rather they will write summaries of different articles+comments. But they will get experience o fboth conditions). Participants will also be asked how easy they found the task/how helpful the tools were in each condition). Responses will be captured either by: i) Likert scale questions which will enable quantification across a set of responses; ii) via questions and comment boxes, which will enable participants to describe other aspects of their experience in the two conditions; or iii) a mix of these two approaches.

2. *A Comment Editor selecting "hot topics" from a set of articles plus comments*

Another task which is currently carried out on a small scale by *The Guardian*, and which was repeatedly mentioned by respondents in our Use Case Questionnaire as something highly desirable, is the identification of "hot topics" in reader comments. Despite its apparent desirability, it is not clear what precisely is meant by "hot topics". For example, "hot topics" could mean news articles that are provoking lots of comment or alternatively issues that are coming up frequently in reader comment across multiple news stories, where these issues may or may not be issues mentioned explicitly in the news article to which the comments are attached. Further refinement of the notion of hot topic in consultation with news professionals and news readers, is required for this evaluation task to be well-defined.

Evaluation setup: participants are given the task of selecting hot topics from a set of news articles with associated comments within some overall time limit (may finish earlier if they wish).

We create two conditions:

- Baseline: participants are given the set of articles with associated comments and a ranking of the articles and/or comments by various criteria such as number of comments/responses, user ratings on the comments, etc. Participants may view the articles and comments in the normal way.
- SENSEI: where they have access to one or more SENSEI functionalities, for example, (1) a Trend Report (Use Case 5) and (2) a Comment Digest (Use Case 6) in addition to the article and comments being displayed in a typical threaded comments interface.

Assessment: Assessment will be carried out using a variety of measures and approaches that will include some or all of:

- time taken to complete the task (ideally maximum time allowed would be varied under the two conditions and the effect on the following measures observed);
- an assessment of the relative quality of hot topic selections made under the two conditions. This could be done via third party assessment using criteria to be determined. Factoring out subject variability is important but could possibly be addressed via a latin square design.
- possibly, assuming feasibility, we will investigate traces by the participants of the process they followed in choosing their set of "hot topics" , e.g. either (a) session logs of participant interaction with the news article, comments and, when available, the Trend Report and Comment Digest or (b) user-generated logs of materials viewed/gathered during the course making their choices (e.g. a short list of candidate hot topics). Here the aim is to explore and ideally quantify differences in user interaction with the source materials in the two conditions in order to understand how SENSEI tools may differently inform the character of the output.
- a post-hoc questionnaire – following completion of the task, participants will be asked to answer a post-hoc questionnaire about their experience. This will ask them to give their perception of how factors such as depth (how deeply they engaged with the comments) and coverage (how broadly they engaged with the comments) differed across the different conditions. Further, we want to investigate how the specific features of the SENSEI functionalities help them to carry out their task: so, in choosing hot topics when provided with a Trend Report and Comment Digest, we would want to know whether the participants believed the selection of content in the Trend Report (which aims to identify important trends in comments in a user-specified time period) and the Comment Digest (which aims to identify the top articles and comments in a time period) helped in making better choices more confidently (note this differs from the previous assessment in not looking directly at what the participants did but in asking them about what they did). Responses will be captured either by Likert-scale questions which will enable quantification across a set of responses or via comment boxes which will enable participants to describe other aspects of their experience in the two conditions. As with the preceding extrinsic evaluation task, participants will be asked to select comments under both system conditions but they will not be asked to do so for the same article plus comments.

3. *A Comment Editor selecting "editor picks" from a set of comments*.

Some on-line news providers, such as The Guardian, manually identify a subset of comments for certain stories as substantively contributing to the debate. The Guardian calls these "Guardian Picks".

Evaluation setup: participants are given the task of selecting editors picks from a set news articles with associated comments within some overall time limit (may finish earlier if they wish).

We create two conditions:

- Baseline: current practice, where they have access to the article and comments in a typical threaded comments interface, including links to commenters' profiles as created currently by the newspaper.
- SENSEI: where they have access to one or more SENSEI functionalities, for example, (1) a SENSEI THM summary (Use Case 1), (2) a Commenter Profile (Use Case 3) and (3) a Comment Digest (Use Case 6) in addition to the article and comments being displayed in a typical threaded comments interface.

Assessment: Assessment will be carried out using a variety of measures and approaches that will include some or all of:

- time taken to complete the task (ideally maximum time allowed would be varied under the two conditions and the effect on the following measures observed);
- an assessment of the relative quality of editor's picks selections made under the two conditions. This could be done via third party assessment using criteria to be determined. Factoring out subject variability is important but could possibly be addressed via a latin square design.
- possibly, assuming feasibility, we will investigate traces by the participants of the process they followed in choosing their set of "editors picks" , e.g. either (a) session logs of participant interaction with the news article, comments and, when available, the THM summary, Commenter Profiles and Comment Cigest or (b) user-generated logs of materials viewed/gathered during the course making their choices (e.g. a short list of candidate comments). Here the aim is to explore and ideally quantify differences in user interaction with the source materials in the two conditions in order to understand how SENSEI tools may differently inform the character of the output.
- a post-hoc questionnaire – following completion of the task, participants will be asked to answer a post-hoc questionnaire about their experience. This will ask them to give their perception of how factors such as depth (how deeply they engaged with the comments) and coverage (how broadly they engaged with the comments) differed across the different conditions. Further, we want to investigate how the specific features of the SENSEI functionalities help them to carry out their task: so, in choosing editor's picks when provided with a THM summary, a Commenter Profile and a Comment Digest, we would want to know whether the participants believed the selection of content in the THM summary (which should identify important significant comments in the comment stream), Commenter Profiles (which should help identify "reasonable" comment providers) and Comment Digest (which includes a set of system-selected top-ranked comments) helped in making better choices more confidently (note this differs from the previous assessment in not looking directly at what the participants did but in asking them about what they did). Responses will be captured either by Likert-scale questions which will enable quantification across a set of responses or via comment boxes which will enable participants to describe other aspects of their experience in the two conditions. As with the preceding extrinsic evaluation task, participants will be asked to

select comments under both system conditions but they will not be asked to do so for the same article plus comments.

4. A comment provider writing a new comment.

Comment providers engage to some degree with a news article and/or existing comments before posting their comment. A question is whether the availability of a combination of SENSEI technologies, e.g. a THM summary (Use Case 1) and/or a Commenter Profile (Use Case 3), would affect the patterns of posting of regular comment providers.

Evaluation setup: participants are given the task of reading a set of news articles and associated comments on a topical issue in the news. We will present participants with a background questionnaire to establish if they have posted on the issue before or if they have an interest in the issue/have been following the issue in the news. After reading each article they are asked determine whether they wish to comment and, if so, whether they wish to start a new comment thread or respond to an existing comment (or both). They are then asked to author their comment.

We create two conditions:

- Baseline: current practice, where they have access to the article and comments in a typical threaded comments interface, including links to commenters' profiles as created currently by the newspaper.
- SENSEI: where they have access to one or more SENSEI technologies, e.g. (1) a SENSEI THM summary (Use Case 1), (2) a Commenter Profile (Use Case 3) and (3) A Similar Content Report (Use Case 4) in addition to the article and comments being displayed in a typical threaded comments interface.

Assessment: Assessment will be carried out using a variety of measures and approaches that will include some or all of:

- time taken to complete the task (ideally maximum time allowed would be varied under the two conditions and the effect on the following measures observed);
- an assessment of the relative quality of the decision to post, positioning of comment and comment content under the two conditions. This could be done via third party assessment using criteria to be determined. Factoring out subject variability is important but could possibly be addressed via a latin square design.
- possibly, assuming feasibility, we will investigate traces by the participants of the process they followed in determining whether and where to post their comment and in authoring the comment , e.g. either (a) session logs of participant interaction with the news article, comments and, when available, the THM summary, commenter profiles and similar content report or (b) user-generated logs of materials viewed/gathered during the course of adding their comment. Here the aim is to explore and ideally quantify differences in user interaction with the source materials in the two conditions in order to understand how SENSEI tools may differently inform the character of the output.

- a post-hoc questionnaire – following completion of the task, participants will be asked to answer a post-hoc questionnaire about their experience. This will ask them to give their perception of how factors such as how easy it was decide whether or not they wished to comment, how easy it was to determine where to comment (new thread or response; response to whom?), and how they managed the process of getting any additional information about the article, topic or commenter they were replying to in the two conditions. Responses will be captured either by Likert-scale questions which will enable quantification across a set of responses or via questions and comment boxes which will enable participants to describe other aspects of their experience in the two conditions.

4.4. Insight Oriented Evaluation

Intrinsic evaluation assesses how well a system performs at the task it was designed to address. Extrinsic evaluation assesses how a system helps a human to complete a task in the larger task environment in which the system is deployed. However, extrinsic evaluation of some new technology is difficult because it may enable activities that are simply not performed at present, perhaps, e.g., because processing the volume of available data is not something that persons or organizations could feasibly undertake. This is true of many new analytics technologies, where the volume of data being analyzed for patterns or insight is simply too huge to have been analysed in the past by humans. Thus, what needs to be assessed in such cases is new behaviours or activities that emerge as a result of the new technologies and/or the provision of items of actionable intelligence, or “insights” that the technology affords.

For example, the availability of a THM summary (Use Case 1) or a Comment Digest (Use Case 6) might lead more people (news professionals or comment readers) to engage with reader comments, people who are currently put off by the sheer number or lack of structure of the comments. A Trend Report (Use Case 5) might help news editors determine which news areas were stimulating most reader interest and hence affect the allocation of scarce news coverage resources to particular areas. An Update and Extend Report (Use Case 2) might help a reporter refine his text for subsequent releases, or suggest angles for new stories.

Evaluation Setup: To assess whether user behavior have changed or whether insights are being afforded by SENSEI tools we propose to identify selected users and give them versions of SENSEI tools to use on their desktop. We will ask them to make an effort to use these tools in the course of their day-to-day interaction with on-line news and reader content over an extended period of time, i.e., ideally several weeks at least. We propose to identify 3 users from each of our user groups (comment readers, comment providers, reporters, sub-editors, comment editors and editors) as participants in this evaluation.

Assessment: Following their period of use of the SENSEI tools participants will be asked to answer a post-hoc questionnaire about their experience. This will ask them to give their perception about whether and how their behavior has changed in relation to how they engage with user comments, given the SENSEI tools. It will also ask them to identify any concrete instances of insights afforded by the tools that led them to do or consider doing something that they would not have done without the tools (mechanisms will be provided to allow them to record these during the trial, so that they do not need to rely on memory at the end of the period). Responses will be captured either by Likert-scale questions, which will enable quantification across a set of responses, or via comment boxes which will enable participants to describe other aspects of their experience.

Ideally we would like to gather session logs for these users with and without the SENSEI tools and explore how their behavior differs. However, it is unlikely this sort of study will be possible during the course of the project given the time consuming nature of such an evaluation and the effort available for this evaluation in the time course of the project. It is also possible that the SENSEI tools may not be robust or user-friendly enough to persuade users to employ them for the sort of longitudinal study proposed above. In this case we shall design a more time-limited study in which users from the news professional and comment reader/provider groups are asked to use the tools to explore a given set of articles and comments with a view to determining, by means of a post-hoc questionnaire or a focus group style interview session, whether they believe the tools afford insights they could not have gained using only existing comment reading and presentation technologies.

4.5. Evaluation Baseline

We report in this Section the baseline values we have been able to identify, with reference to our use cases, and we will explain why for insight – oriented evaluation the baseline values need to be understood differently.

For synopsis generation of call-centre conversations, the following extractive summarization baselines will be used. Each method selects a subset of speech turns from the conversation; the length limit is 7% of the number of words in the conversation:

- LB: Longest speech turns within the first quarter of the conversation, supposedly capturing the problem statement from the caller. Using the few first turns of the conversation would not be relevant (as it is in text summarization) because they often contain the opening, made of courtesy formulas and general inqueries.
- LA: Longest speech turns of the conversation.
- LE: Longest speech turns in the last quarter of the conversation, supposedly capturing the solution to the callers' problem.
- MMR: Maximal Marginal Relevance (Carbonnell et al., 1998).
- RS: Random selection of turns, repeated 100 times.

In addition to these baselines, high-performance baselines will be considered by implementing and adapting successful text summarization systems, such as those evaluated in (Hong et al, LREC 2014).

For ACOF filling, the system must generate "yes/no/na" labels for each question and is evaluated in term of accuracy. The following baselines will be evaluated:

- YES: answer "yes" to all questions.
- NO: answer "no" to all questions.
- NA: answer "not applicable" to all questions.
- MF: the most frequent answer for each question in the training data.

- RND: random answer, repeated 100 times.
- NB: a naive bayes model which outputs the answer with the highest product of P (answer|word) for all words in the conversation, P being estimated with maximum likelihood with Laplace smoothing.

5. Conclusions and Further Work

This deliverable illustrates the progresses related with 1) the definition of SENSEI Use Cases, and 2) the evaluation methodology that the Consortium will apply to assess performances and usefulness of the SENSEI technologies. It builds upon the content of D1.1, where a range of possible scenarios of use where described for applying SENSEI conversational summarization technologies in both speech and social media. While D1.1 scenarios were already been identified by interacting with potential users, the selection, refinement, and re-structuring of the D1.2 use cases takes into account the results of applying qualitative research methodologies such as focus groups, user surveys, and interviews, in the requirement collection research phase. The result of this activity is a selection and prioritization of the original set of use cases.

D1.2 also presents the SENSEI evaluation framework, which is organized into a tripartite model including metrics that are applied at the technology level and at the prototype level. The first set of metrics includes state of the art evaluation tools and metrics, which will be applied throughout the project to monitor the performances of the summarization and discourse technologies. At the prototype level we envisage that the SENSEI prototype will be evaluated on the basis of intrinsic, extrinsic, and insight-oriented evaluation settings. In D1.2 the related evaluation tasks have been described for both domains. In this process we have been trying to deal with the problems posed by the novelty of the tasks that users will be able to accomplish by using SENSEI technologies: if the task is completely new, such as in case of application of summarization in call centres and in social media, no well-established baselines and evaluation metrics exist in the literature. We will deal with this problem by designing analyses and collecting insights from users, as described in the section devoted to insight-oriented evaluation.

Finally, D1.2 is accompanied by a set of Appendices where the interview questions, focus group materials, and guidelines are reported.

Appendix A: Speech Use Case Qualitative User Data

In this Appendix we report the materials used in the two focus groups and semi-structured interview we held on April, 2nd and July 16th with the final users of the speech scenarios, i.e. with the Coordinator of the Front Office Supervisors, with the Quality Director, and Front Office Agents' Supervisors of TP.

The two focus groups were moderated by a licensed psychologist who also took notes during the discussion.

The first table below reports the questions used by the Moderator to facilitate the discussion. The Introductory part of both focus groups included the presentation of the SENSEI goals (first focus group), and the presentation of the a draft ACOF that had been used by two of the participants for evaluating call centre agents performance over real call centre calls (second focus group).

Questions used during the first focus group

This focus group had four participants, plus the Moderator, It was held when we were designing the speech use cases reported in D1.1. The goal was to collect users' interest with respect to SENSEI technologies, understanding if and at what extent they could be applied in a call centre working environment, which activities of the agents' supervisors they could have impact on. The background of the focus group discussion was constituted by the description of Quality Assurance activities provided by TP partner during M1-M4 of the project. That detailed description was reported in D1.1.

Table 3 – First Focus Group: questions used by the Moderator

Introduction	(1) Welcome (2) Overview of the topic (3) Ground rules
Discussion Questions	Five Types of Questions • Introductory Question (IQ) • Transition Questions (TQ) • Key Questions (KQ)
Ending Questions	Three Types of Questions • All things considered questions (ATCQ) • Summary question (SQ) • Final question (FQ)
Questions asked in the SENSEI focus groups	**Welcome:** Thank you for the time you are dedicating to take part in this focus group. You have been invited because of your role in the organization of your call centre company. **Overview of the topic**

	In SENSEI we are developing new functionalities that we believe will help users to access monitor rapidly and effectively some of the events happening in agent-customer conversations going on in call centre. We need to develop use cases as part of this work, to ensure the technologies we will design will be based on real user needs. **Ground rules:** I will ask you some questions about your daily job, and my role is Moderator. I will take notes while you're speaking. In the discussion you don't need to agree with others' opinions, but just report your experience: there are no right or wrong answers, maybe only differing points of view about how much valuable can be the technologies we are going to develop in SENSEI. **Questions (their types in brackets):** [Note: TQs were not used in this focus group] "Please describe your role in the agent monitoring process" (IQ) "From a previous overview of your daily activities, we could learn that during your call listening task you rely on a set of items pointing at agents' behavior you may want to focus ("scheda d'ascolto"). Can you describe how do you select the relevant items?" (KQ) "Which aspects of the agent behavior are most important for filling the listening forms?" (KQ) "How many calls do you need to listen to for filling one form?" (KQ) "What do you think about the possibility of relying on technologies that could support your listening activities?" (KQ) "Should the listening form be generated automatically, do you think that could simplify your task?" (KQ) "Should you have summaries of the conversations going on in your call centre, can you imagine how you could use them for your job?" (KQ)

	“Which activities included in the listening process could benefit from automatically generated call summaries?” (KQ, asked if the previous question gets positive answers) "Of all the things we discussed, what to you think is the most important?" (ATCQ) "Is this an adequate summary?" (SQ) "Have we missed anything?" (FQ)
Notes	
Quotes	Quotes will be made available on requests due to confidentiality issues.
Key Points / Themes	✓ Different kinds of listening to agent – customer conversations ✓ Generation process of the listening forms ✓ Agent behavior in focus: speech, tone of voice, lexical choices, compliance with internal procedures. ✓ Nr of calls listened for filling a single agent listening form (on average): 4; this is a rough estimation, because it depends on the complexity of the agent task, and length of the calls ✓ Positive expectations about the technologies ✓ Would like to have online automatic monitoring of call centre calls ✓ Would like to have automatic detection of user satisfaction ✓ The availability of automatically generated listening forms could improve their work with respects to the opportunity of focusing on most problematic calls or call segments while performing human evaluation ✓ They imagine benefit in terms of time-to-completion of the listening task ✓ Summarization technologies would be something completely new ✓ They can imagine using summaries for quickly detecting topic of the conversations, especially in the outbound and customer care campaigns.

Questions used during the second focus group

The second focus group was held after that the Consortium selected the automatic generation of ACOFs and synopses generation as priority use cases to be developed in SENSEI. TP and UNITN elaborated simplified ACOFs that were submitted to call centre QA supervisors. They were asked to use that drafted ACOF, and to generate sample synopsis by following AMU Guidelines reported in the next section. The focus group goal was to collect their impression based on those experiences, and to refine the use cases.

Table 4 – Second Focus Group: questions used by the Moderator

Introduction	(1) Welcome (2) Overview of the topic (3) Ground rules
Discussion Questions	Five Types of Questions • Introductory Question (IQ) • Transition Questions (TQ) • Key Questions (KQ)
Ending Questions	Three Types of Questions • All things considered questions (ATCQ) • Summary question (SQ) • Final question (FQ)
Questions asked in the SENSEI focus groups	**Welcome:** Thank you for the time you are dedicating to take part in this focus group. You have been invited because of your kind availability in testing on real conversations the listening form we have provided, and for writing the call summaries. **Overview of the topic** We want to collect your impressions, and recommendations based on your experience while performing such tasks. **Ground rules:** I will ask you some questions as in the previous meeting we had in April, and my role is acting as Moderator. I will take notes while you’re speaking. In the discussion you don’t need to agree with others’ opinions, but just report your experience. We will also review the summaries you wrote: there are no right or wrong summaries, as you could learn from the guidelines, but only differing points of view and different ways of phrasing the content you wanted to report.

	Questions (their types in brackets): [Note: TQs were not used in this focus group] “Please describe if and at what extent you were ease while using the listening forms provided to you” (IQ) “Could you feel in the form all the relevant aspects of agent behavior you needed to assess his/her performance?” (KQ) “Were you able to complete the listening forms on the basis of listening to a single call?” (KQ) “Do you think that the listening form can be modified, and if yes, what do you suggest?” (KQ) “This experience has modified your previous opinion about the benefit of relying on technologies that could support your listening activities?” (KQ) “Did you feel comfortable with the summary generation guidelines provided to you?” (KQ) “Now that you know what a synopses is, can you imagine some possible use of them in your Quality Assurance process?” (KQ) "Of all the things we discussed, what do you think is the most important?" (ATCQ) "Is this an adequate summary?" (SQ) "Have we missed anything?" (FQ)
Notes	
Quotes	Quotes will be made available on requests due to confidentiality issues.
Key Points / Themes	✓ The listening form provided to them was easy to use ✓ Since it abstracted from idiosyncrasies related with technical tasks of the agent, a single form could be filled most of the time by listening to a single conversation ✓ The form covers all the items usually needed for evaluating agent behavior

	✓ They would like to compute a score starting from the agent observation form ✓ They propose Agent Observation Form (ACOFs) as the name of the listening form ✓ Need to provide more training for synopses annotation ✓ It would be interesting to relate topics of the calls as they emerge from synopsis with ACOFs

Appendix B: The SENSEI ACOF

Table 5 below reproduces the SENSEI call survey grid (Agent Observation Form). We provide brief comments for each group of evaluation items.

Table 5 – SENSEI call survey grid (ACOF)

SENSEI ACOF				
	YES	NO	NA	Notes
Call Opening				
Agent complies with call opening protocol				
Problem Identification				
Agent's listening is focused on Client's problem and his/her questions are relevant				
Problem Fixing				
Agent is competent with respect to the product and applications				
Agent explanations are clear, complete and simple				
Agent applies the correct protocol				
Increasing Value				
Agent can manage Client's objections, reassure him/her, and takes in great consideration Client's satisfaction				
Summary, Confirmation, and Call Closing				
Agent is self-confident				
Agent speech is positively connotated				

Agent complies with the call closing protocol				
Communication Skills				
Agent is polite but purposeful with the Client				
While being professional, Agent is able to adequate his/her speech to the Client's speaking style				
Time management: Agent manage the possible waiting times by explaining their rationale				
Listening capability				

The ACOF is organized into six main areas. Two of them, the first area and the last one, take into account the conversational style of the agents, i.e. their abilities in opening and closing the calls appropriately, and communicate effectively. Also the evaluation item concerning time management aims to check at what extent the agent is able to communicate with the customer when there are possible delays due to operations that s/he needs to perform on the system. An important issue is the agent listening capability: the Front Office Supervisor evaluates the conversational attitudes of the agents concerning turn taking, providing feedback, etc. It is expected that agents rated with good values for listening capabilities are able to run calls where empathic behaviour can be assessed.

Also the third item of the "Summary, Confirmation, and Call Closing Area", i.e. "Agent complies with the call closing protocol" might include the evaluation of communicative skills, but what is under examination here is more the ability of the agent of applying some well defined protocol of call closing, that may vary across different selling or customer care campaign, but that always include exact summarization of the actions or suggestions provided in the call.

Problem Identification, Problem Fixing, and Increasing Value, i.e. the three central areas of the ACOF, are more focused on the ability of the agents in managing call content. In other words, they are more focused on the call management and call resolution. It is likely that high values in these three areas may correlate with high value of call resolution in the call centre monitoring statistics. In addition, low values assigned to these items may be used for identifying possible areas of need for further training of the agents.

In the Notes column for each item the Front Office Supervisor may add comments, i.e. sentences that motivate the value assigned. In the large ACOF database provided by TP we could observe that the Notes are always provided by the evaluators when the assessment is negative.

Appendix C: Guidelines for Call Summaries

The Guidelines that have been provided by AMU, and used by TP and AMU for the summary/synopses annotation of the LUNA and DECODA corpora, are reported below.

Summary / Synopsis Guide
(DECODA corpus inspired version)

Why do we need this document?

Making a summary is a very subjective task. Two different persons won' t systematically pick up the same information, or the same level of detail, and then won't write the same summary.

Our systems need these human summary to learn how to automatically generate them in a good way, and to be evaluated as well. Here we there are two options:

The first option is to hire a lot of annotator and let them write their summaries in a free way. And then make a study on all the resource collected to find a good shape for a summary. But this method is very expensive.

The second option is to write a guide to lead the annotators in a certain way of annotation to get some similarity in the summaries generated. In this way we introduce some guideline depending on the corpus to be annotated (previously studied). The annotators are not free anymore, but then only a few can annotate the corpus. Because of that kind of guide, all the summaries generated will come with some similarity and become more usable for our systems.

First of all, let’ s define here what we call a synopsis. A synopsis is a summary of the call taken from a call centre. This call mostly involved two persons, the caller (here someone who wants something from the call centre) and the adviser (the one who answer in the call centre and give some solution to the caller).

After some studies we concluded that this synopsis should be no longer than 7% of the original call (in terms of words). We are aware that the annotator won’ t spend their time counting the words··· So we probably include this length limit in the next version of the interface. Speaking of the interface, right now, it’ s just a simple and minimalist interface that just provide the conversation (spoken and written) and a box to fill with the synopsis. (Everything should be revamp to a better version).

Go back to the synopsis, this is a pretty subjective things to do, that’ s why we’ ll try in this guide to give some line to follow.

First we can distinguish two kind of synopsis:

1. The semantic oriented synopsis: that focused more on the content of the conversation than the end of it.
2. The “structure” oriented synopsis: That focused more on the way the adviser treat the caller
and his call.

For the rest of this guide, only the semantic oriented synopsis we’ll be considered (the “structure” oriented synopsis could be developed in another guide).

Syntax, length, way to write

You have to keep in mind that you are limited in term of length. Moreover we don’t really need to get some really good language, what we need the most is the information!

Here are some tips:

- Only pick the important information in the call.
- The shorter your sentence is the better your synopsis will be
- Do not hesitate to make a “telegraphic style” sentence (e.g. “Route request in Paris centre”)
- Try to sum the long exchange with a simple action, or if there isn’ t any good information in it don’ t even pick it up in the summary.

Practical cases

Nothing’ s better than a good example. Here are three of them translated from French with some comments on how we ended here:

Example 1

- Hello - Hello - Hello - Eh it's Mrs [*name removed*] eh I was calling you because I lost my scarf, where hmm in the bus hmm 140, when - Yes - I left at Colombes [*name*] yesterday night between - Yes you need to call later madam, around 11am, it's too early, it's not open yet. - At 11 - 11am yes, ok? - Ok - See you later.
Synopsis 1: Needs information about a scarf lost in the bus 140. Wait the service opening and call later.
Synopsis 2: / Maybe there is an error made by the annotator here about the bag / Bag lost in the bus 140, but call at 11am when the service will be open.

Comments:

Here's the call is pretty clear, the caller just want to know if there is any news about her loss.

Then both annotators ended with the result of the call (e.g. call later when the service is opened).

We could argue a bit on it due to the length of the synopsis, but it's a pretty important information and the length of the original conversation is pretty short.

Example 2

- Hello
- please
- hello?
- Yes, hello
- hello
- I call you because I would like to have some information, when I'm at the Javel's station
- yes
- of the RER [*the name of the train*], is there a bus that could drive me closer to the Vauthier's
- street at Boulogne [name of the city] without taking the metro
- Vauthier's street at BoulogneBillancourt
- yes
- just a moment, I'm looking for it
- thanks
- Madam
- yes
- you have the 72, you cross the Mirabeau's bridge to the Mirabeau's stop in way to SaintCloud's park.
- yes
- and you stop at the Reine JeanJauras'road's stop
- Wait, at the end of the Mirabeau's bridge I take the 72
- 72 to SaintCloud and you stop at the Reine JeanJaures'road's stop
- Ok
- Ok thank you
- Good bye
- Good bye
- Thank you

Synopsis 1:
Needs for a connection by bus between the exit of the RER and a street in Boulogne.

Synopsis 2:
Needs for a connection by bus between the RER and the Vaulthier's street at Boulogne.

Synopsis 3:
Information on a potential bus connecting the Javel's station of the RER and a street in Boulogne.

Comments:

As you can see here, the three annotators picked up the same information, only the syntax is slightly different. Some annotators are more accurate about the name of the street (i.e Vaulthier's street / street in Boulogne).

Example 3

- Hello
- Yes
- hello madam, eh I'd like to have some information
- Yes
- I'd like to know if the bus centre strike at Vitry [*city*] will continue tomorrow or if it was just today?
- No it's, it's just today, tomorrow everything is normal
- Ok thank you very much
- My pleasure
- goodbye
- Have a good day, goodbye

Synopsis 1:
Needs information about the renewal of the strike. Normal traffic in prevision.

Synopsis 2:
Renewal of the bus centre strike at Vitry [*city*], no tomorrow the traffic is normal.

Comments:

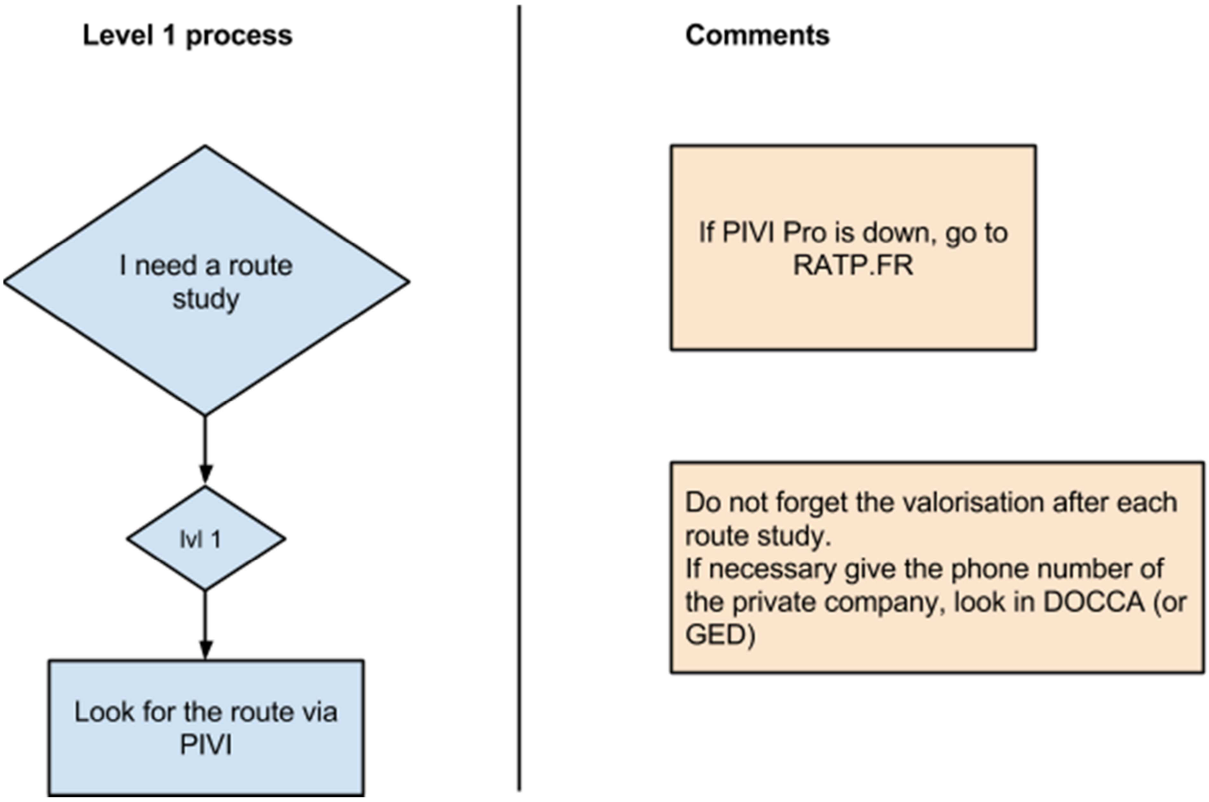
Nothing special to mention here.

However we can notice the syntax used by annotator 1. Since the beginning he used the same method i.e. “Needs information about” . It can be a bit “word consuming” but it can also be a good way to introduce some kind of structure in every synopsis.

Call scenario

The DECODA's conversation come from a call centre from the RATP. In this call centre every adviser has some call scenario to help them answering the caller.

For example, the caller needs to know a route, then the adviser get his route scenario and answer the question. The call scenario looks like this:



If PIVI Pro is down, go to RATP.FR

Do not forget the valorisation after each route study.
If necessary give the phone number of the private company, look in DOCCA (or GED)

The call scenario is divided in two parts, the call process and the comments.

The comments are just some additional stuff to help the adviser during the process, it's generally very technical and then not useful for us. On the other hand the call process is very interesting.

The call process is like a state diagram. On the top on the diagram we have the main issue of the caller (here "I need a route study"). Then by following the arrow we have the way to answer depending on the other caller's issues.

Usually these call scenario draw a good base for a synopsis because if you can apply the corresponding scenario to a conversation it will give you the main theme, and the process pretty much already summed up.

How to get or generate these call scenario?

In the DECODA corpus we already got these call scenario because of the nature of the corpus.

But how could we make them if they are not available for other conversations? Basically the first thing to do is to establish a list with all the main theme of the conversation.

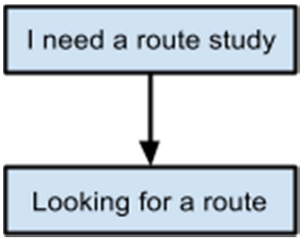
For example in DECODA we have:

1. Route
2. Loses, theft, found

3. Official report
4. Reimbursement
5. Delay, incident
6. Prices
7. Accidents

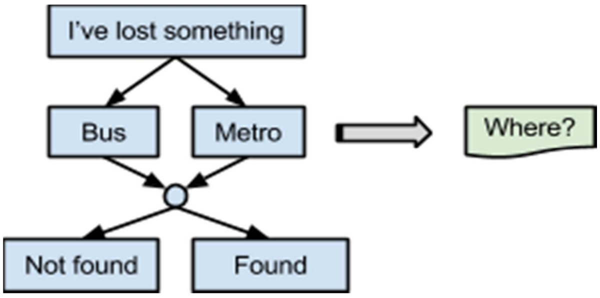
Then for each theme you have to find if there is some redundant schema.

For instance for a route study, in most cases the caller is asking for a route and then the adviser is giving the caller the route needed. It can easily be drawn like this:



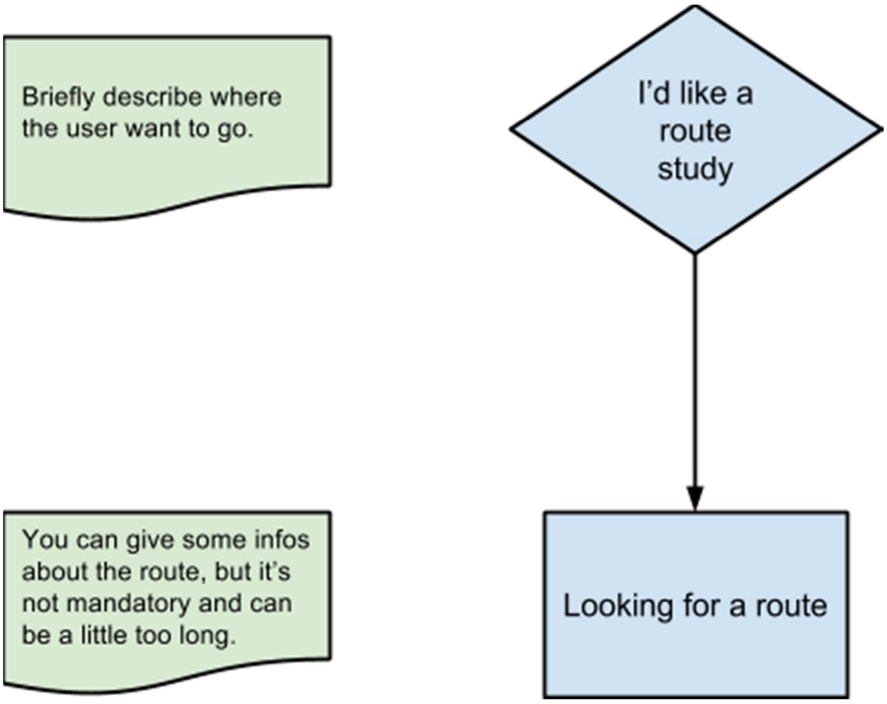
While the state diagram can provide a lot of options, you can add some variety in it.

For instance in case of a lost, we can have several options like where the object has been lost, or is it found yet or not.



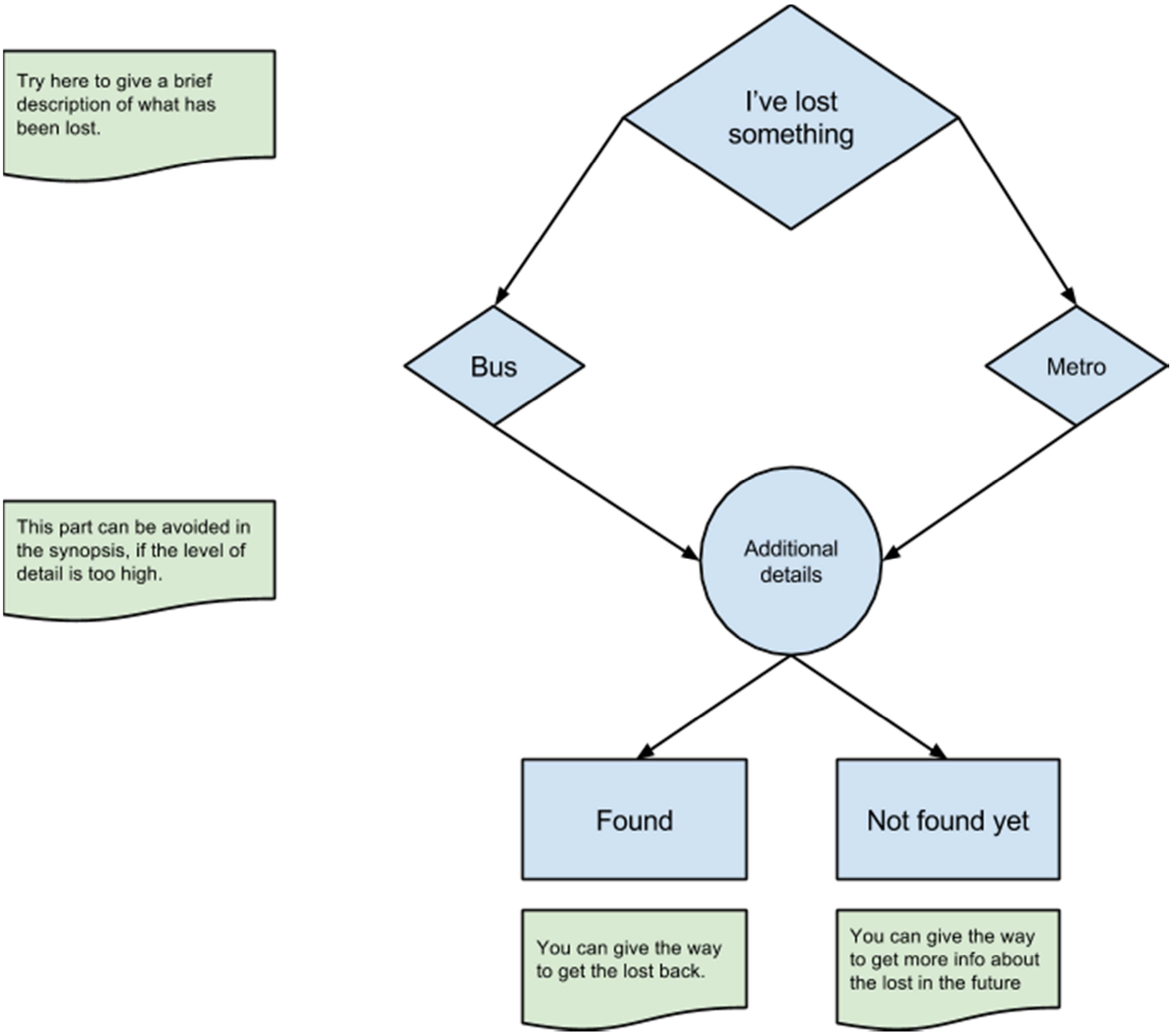
The next part gather the call scenario from DECODA listed as example above. Some comments have been added about the summarization task.

1. Route


Briefly describe where the user want to go.
I'd like a route study
You can give some infos about the route, but it's not mandatory and can be a little too long.
Looking for a route

In case of several route requests in the same call, try not to focus on the destination, try to find a location where all the routes happen like for example Paris centre if there are 2 or 3 route in individual different locations in Paris centre.

2. Loses, theft, found



Try here to give a brief description of what has been lost.

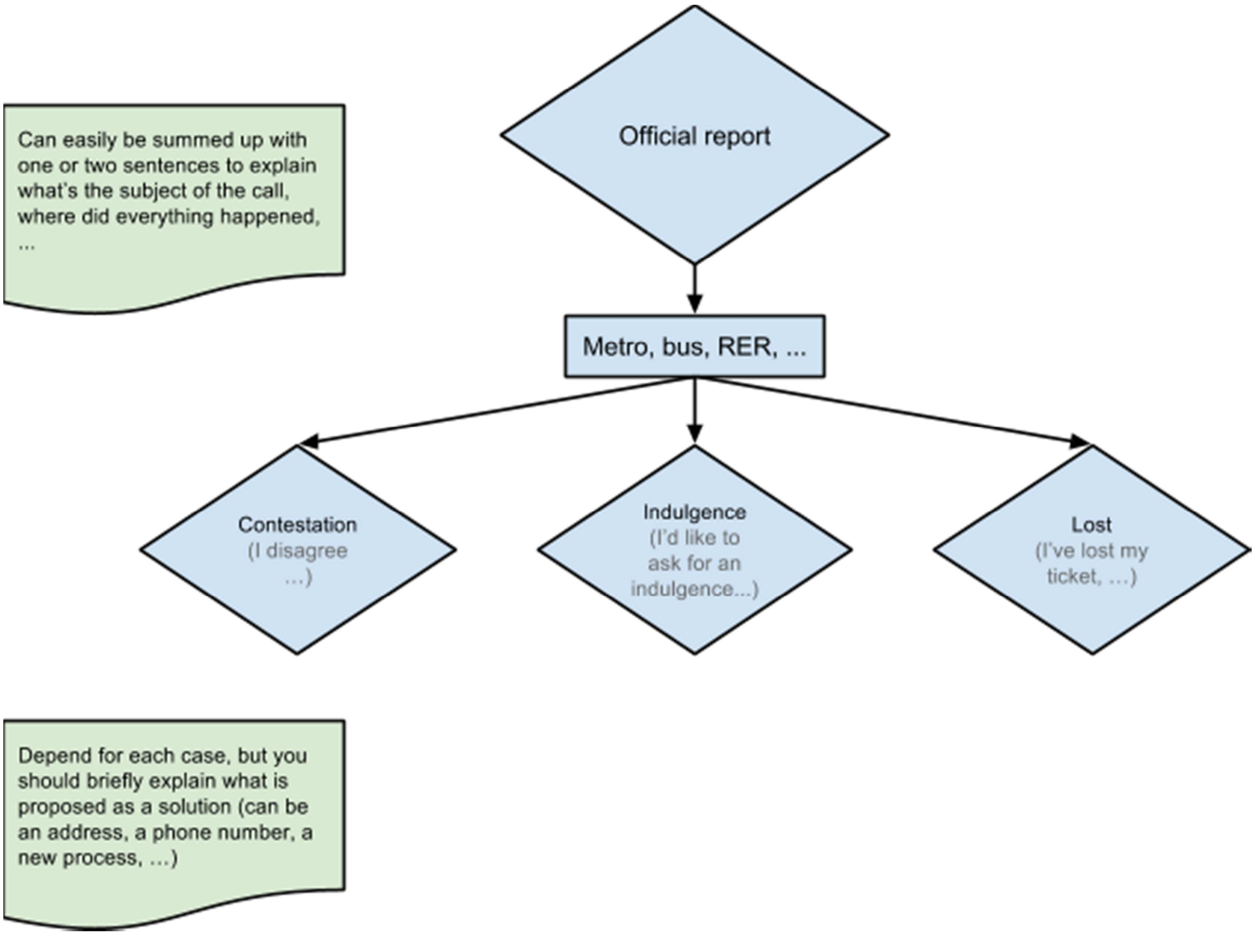
This part can be avoided in the synopsis, if the level of detail is too high.

You can give the way to get the lost back.

You can give the way to get more info about the lost in the future

In this particular case we like to precise if the object has been found or not just by adding at the end “object found, go to [location] to get it back” or “object not found still [recall later]”.

3 Official report

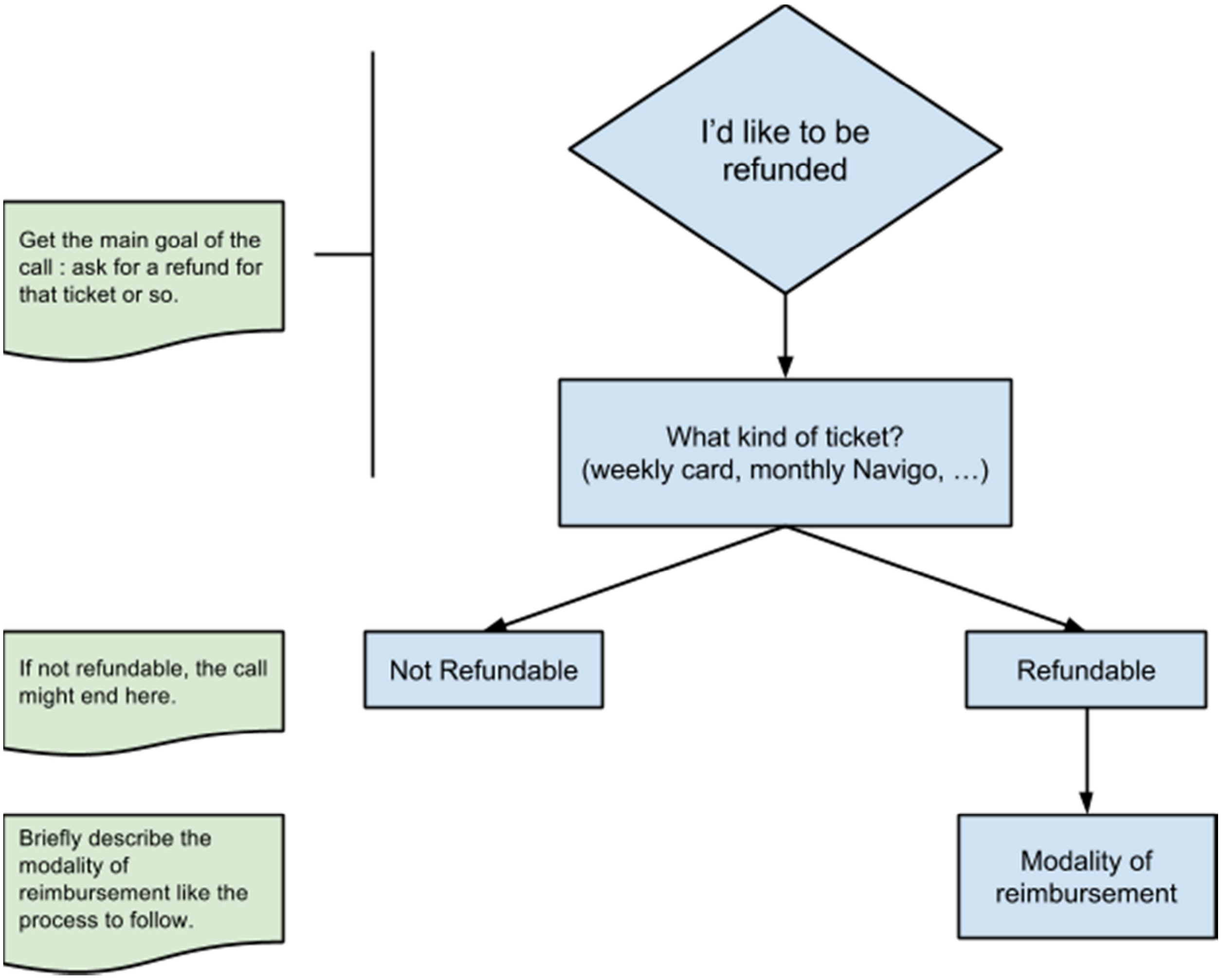


Can easily be summed up with one or two sentences to explain what's the subject of the call, where did everything happened, ...

Depend for each case, but you should briefly explain what is proposed as a solution (can be an address, a phone number, a new process, ...)

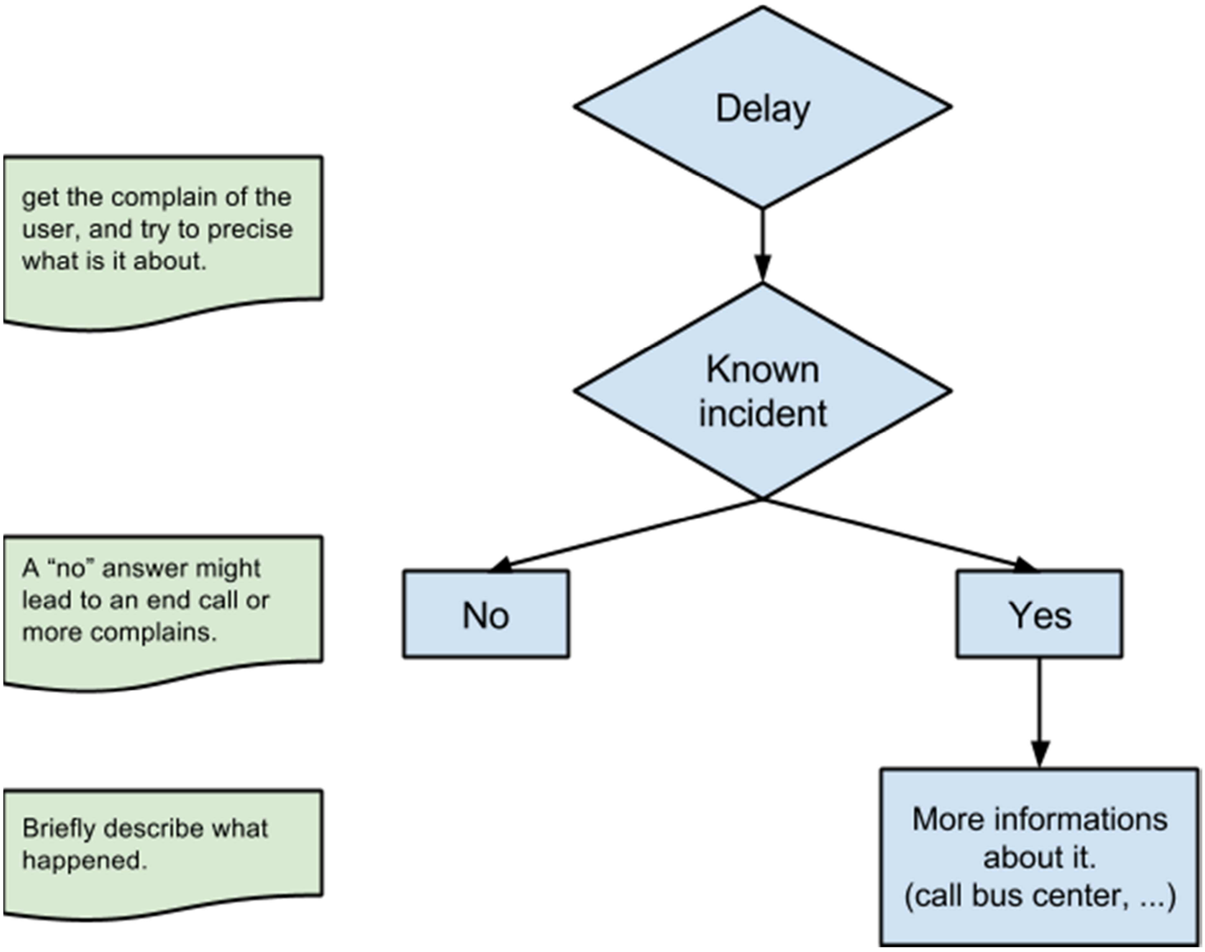
Official report calls are frequently pretty long (to explain everything or so). Basically we try here to sum up the report in one or two sentences and then briefly give the solution given by the adviser (like “communication of the right service”).

4 Reimbursements


Get the main goal of the call : ask for a refund for that ticket or so.
I'd like to be refunded
What kind of ticket? (weekly card, monthly Navigo, …)
If not refundable, the call might end here.
Not Refundable
Refundable
Briefly describe the modality of reimbursement like the process to follow.
Modality of reimbursement

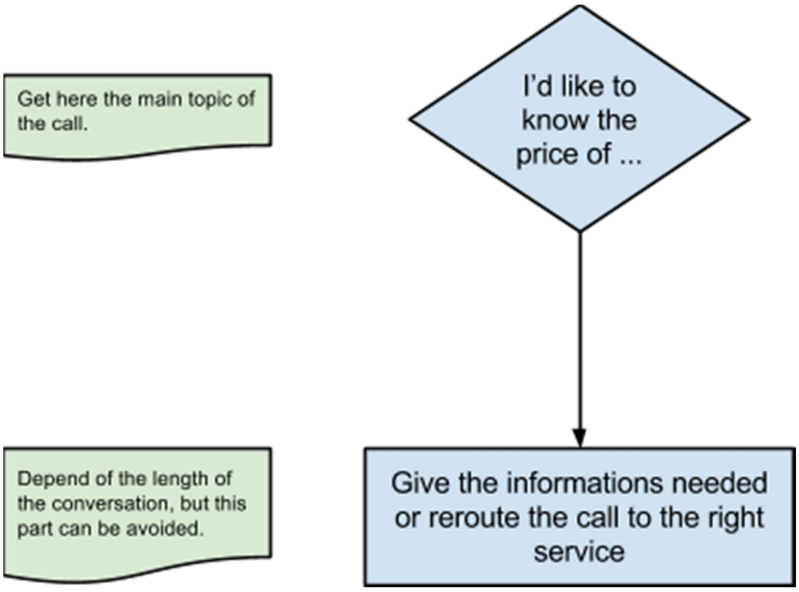
There is a lot of misunderstanding in this kind of call, depending on the knowledge of the caller, but because of the variety of the cards/contracts/errors/··· the exchange between the caller and the adviser are pretty numerous. So just get the main reimbursement issue summed up and then as for the other scenario get the solution if it's relevant.

5 Delay, incident


Delay
get the complain of the user, and try to precise what is it about.
Known incident
A "no" answer might lead to an end call or more complains.
No
Yes
Briefly describe what happened.
More informations about it. (call bus center, ...)

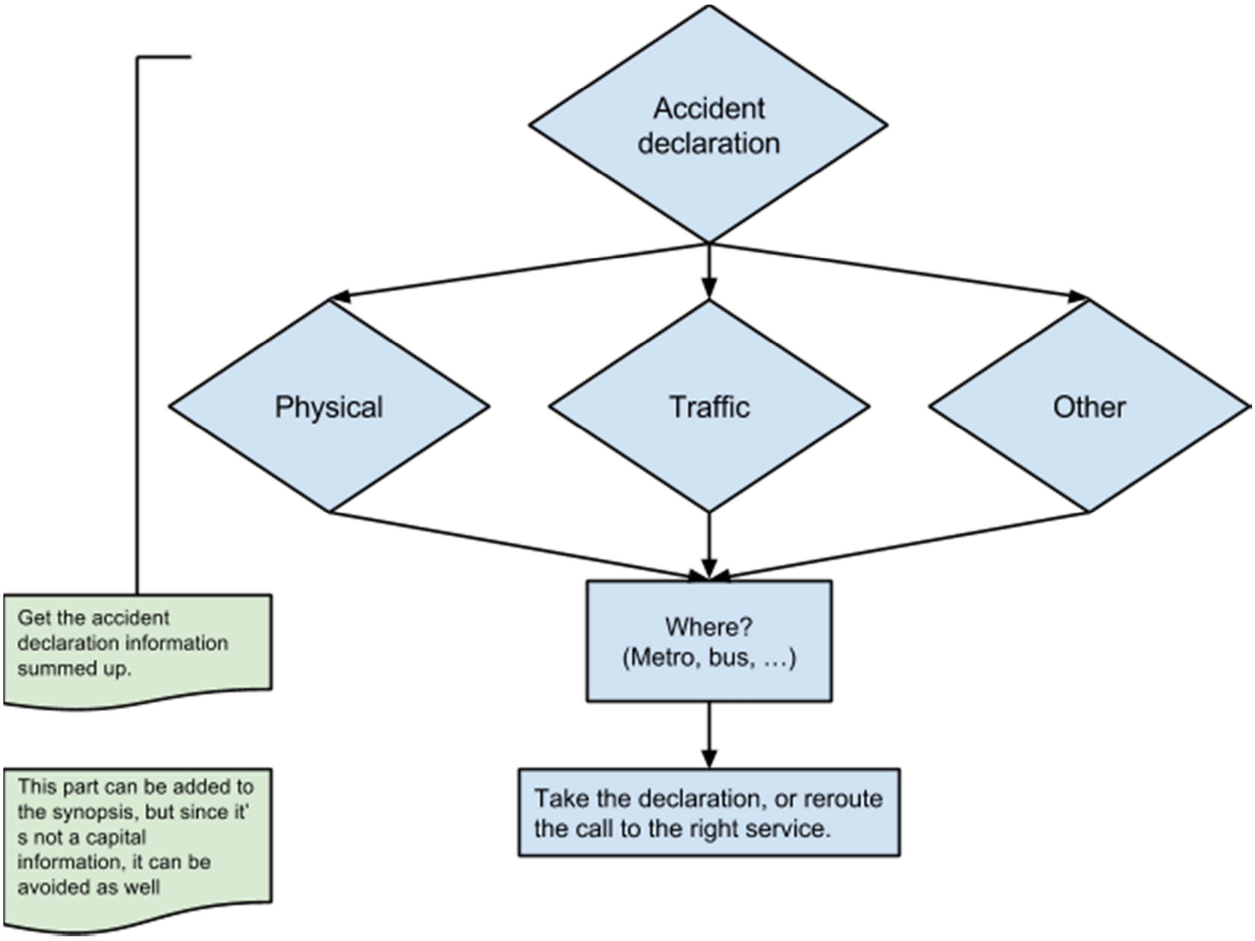
In these calls the adviser will frequently call the bus centre or another service to ask for more information. Generally we don't really consider that call in call, but we prefer to sum up the whole thing just with the final answer/solution to give to the caller.

6 Prices


Get here the main topic of the call.
I'd like to know the price of ...
Depend of the length of the conversation, but this part can be avoided.
Give the informations needed or reroute the call to the right service

Previous scenario (5 Delay, incident). In some cases (the easiest and the shortest) we like to just sum up the whole conversation by the description of the issue (e.g. "asking for prices about an orange card").

7 Accidents


Accident declaration
Physical
Traffic
Other
Where? (Metro, bus, …)
Take the declaration, or reroute the call to the right service.
Get the accident declaration information summed up.
This part can be added to the synopsis, but since it' s not a capital information, it can be avoided as well

Pretty much the same as 5 Delay, incident. We sum up the declaration, and then get the solution/answer.

Appendix D: Questionnaires and Interview Data for Social Media Use Cases

We have designed two types of questionnaires, one to gather background information on the respondents and one to gather feedback from the respondents on the use cases. To gather the background information on the respondents, we have designed two different versions for the two user groups: **news-media professionals** and **members of the public** (i.e. readers of on-line news and comment and comment providers). We show both versions in this Appendix.

We also designed two variants of the use case questionnaires for these two different user groups. To save space, we present only the questionnaires shown for news-media professionals. Furthermore, in the versions of the use case questionnaires presented to the participants, the full text of each of the use cases, as presented above in Sections 3.1, is included at the very start of the section for each use case. Again, this is omitted here to save space. A demo version for the news-media professionals can be viewed at: http://paramita.staff.shef.ac.uk/sensei/demo/introduction.php.

News Media Professionals: Participant Profile Questions

Before proceeding to the questionnaire, we would be grateful if you could tell us a little about yourself and your experience using on-line news and reader comments.

Professional Experience of On-line News and Comment
Please answer the following questions on the basis of your use of on-line news and comment in a **professional capacity**, i.e. where you use the news and/or comments as part of your work.

1. a) Which of the following best describes your current role as a news-media professional?

    ○ Reporter
    ○ Sub-editor
    ○ Comment Editor
    ○ Comment Moderator
    ○ Executive/Editorial
    ○ Educator/Academic
    ○ Other (please specify): ___________________________

    b) Please indicate any previous roles you have experience of in the news-media. (Please select **one or more** options from the list below.)

    ☑ Reporter
    ☑ Sub-editor
    ☑ Comment Editor
    ☑ Comment Moderator
    ☑ Executive/Editorial

☑ Educator/Academic
☑ Other (please specify): _________________________

2. In your current role as a news-media professional, please tell us how often you engage with (e.g. read or skim) any on-line news web sites, e.g. The Guardian, BBC, The Independent, Mirror, etc.?
   ○ At least once a day
   ○ At least once a week
   ○ At least once a month
   ○ Very rarely (i.e. more than one month intervals between visits)
   ○ Never

3. In your current role as a news-media professional, approximately how many distinct on-line news articles do you engage with (read or skim etc.) in a typical day?
   ○ None- I rarely read the on-line news as part of my work
   ○ 1-5
   ○ 6-10
   ○ More than 10
   ○ Don't know

4. In your experience as a news-media professional, when (if ever) did you first engage with (e.g. read or browse or post) **reader comments** in on-line news websites? (Please select the option that best describes your experience.)
   ○ Within the past month
   ○ Within the past year
   ○ Within the past 2 years
   ○ More than 2 years ago
   ○ Never

5. In your current role as a news-media professional, please tell us how often you engage with the **reader comments** in on-line news web sites? (Please select the option that best describes your experience.)
   ○ At least once a day
   ○ At least once a week
   ○ At least once a month
   ○ Very rarely (i.e. more than one month intervals between visits)
   ○ Never

6. a) In your current role as a news-media professional, when you engage with reader comments how many individual comments do you typically read?
   ○ 1-5
   ○ 6-10
   ○ 11-20
   ○ 21-50
   ○ More than 50
   ○ Don't know
   ○ Question does not apply to me – I don't engage with reader comments

   b) If you engage with (e.g. read or skim) comments in on-line news, as part of your work, please tell us a bit about why you do so? (*optional*)

(e.g. "I am a comment editor and I look for interesting comments to recommend"; "I like to see what people have to say in response to something I have written"; etc.)

c) If you *do not* engage with (e.g. read or skim) comments in on-line news, as part of your work, please tell us a bit about why you do not? (*optional*)
(e.g. "I don't have time"; etc.)

7. If in your role as a news-media professional, you are directly involved in producing content for on-line news, e.g. writing or editing copy, for what proportion of the articles that you contribute to do you read the associated comments?
   - ○ 1- 25%
   - ○ 26-50%
   - ○ 51-75%
   - ○ 76-100%
   - ○ Question does not apply to me

8. In your current role as a news-media professional, how often do you post comments? (Please select the option that most accurately describes your experience.)
   - ○ At least once per day
   - ○ At least once per week
   - ○ At least once per month
   - ○ Very rarely (i.e. more than one month intervals between posts)
   - ○ Never

9. In your current role as a news-media professional, if you post comments on a regular basis then approximately how many posts do you typically make in a day:
   - ○ 1-5
   - ○ 6-10
   - ○ 11-20
   - ○ More than 20
   - ○ Don’t know
   - ○ Question does not apply to me

10. a) Please tell us a bit about how you post comment as part of your work. (Please select **one or more** options from the list below.)
   - ☑ I never post comments
   - ☑ I often post a comment in response to something I have read in an article
   - ☑ I often post a comment in response to something I read in the comments
   - ☑ When posting a comment I typically begin a new thread
   - ☑ When posting a comment I only occasionally begin a new thread
   - ☑ When posting a comment, I typically add the comment to an existing thread
   - ☑ When posting a comment, I only occasionally add the comment to an existing thread
   - ☑ If someone posts a reply to one of my comments, I rarely post a reply back to them

☑ If someone posts a reply to one of my comments, I often post a reply back to them
☑ None of the above

b) If you post comments as part of your work, please tell us a bit about how and why you post comments? (*optional*)
(e.g. "I post comments to correct inaccuracies I see in the article"; "I post comments on topics I am interested in"; etc.)

e) If you *do not* post comments as part of your work, please tell us a bit about why you do not? (*optional*)
(e.g. "I don't have time"; etc.)

Other Experience of On-line News and Comment
Please answer the following questions on the basis of your use of on-line news and comment in **a non-professional capacity**, i.e. where you use the news and/or comments for personal interest, and not as part of your work.

11. Please tell us how often you engage with (e.g. read or skim) on-line news web-sites, for personal interest?
(Please select the option that best describes your experience.)
○ At least once a day
○ At least once a week
○ At least once a month
○ Very rarely (i.e. more than one month intervals between visits)
○ Never

12. Please tell us how often you engage with **reader comments** in on-line news web-sites, for personal interest?
(Please select the option that best describes your experience.)
○ At least once a day
○ At least once a week
○ Monthly-At least once a month
○ Very rarely (i.e. more than one month intervals between visits)
○ Never

13. Please tell us how often you post reader comments in on-line news web sites, for personal interest?
○ At least once per day
○ At least once per week
○ At least once per month
○ Very rarely (i.e. more than one month intervals between posts)
○ Never

Thank you, you have completed the background questions. Please proceed to the main questionnaire.

Public Users of On-line News and Comment: Participant Profile Questions

Before proceeding to the questionnaire, we would be grateful if you could tell us a little about yourself and your experience using on-line news and reader comments.

1. Which of the following best describes you?

   ○ **News and comment reader** - I read news and/or comments but very rarely post comments

   ○ **Comment provider** - I read news and/or comments and provide/post comments on a regular basis

2. Please tell us how often you engage with (e.g. read or skim) any on-line news web sites, e.g. The Guardian, BBC, The Independent, Mirror, etc.?

   ○ At least once a day
   ○ At least once a week
   ○ At least once a month
   ○ Very rarely (i.e. more than one month intervals between visits)
   ○ Never

3. Approximately how many distinct on-line news articles do you engage with (e.g. read or skim) in a typical day?

   ○ None- I rarely read the on-line news as part of my work
   ○ 1-5
   ○ 6-10
   ○ More than 10
   ○ Don’t know

4. In your experience, when (if ever) did you first engage with (e.g. read or browse or post) **reader comments** in on-line news websites?
   (Please select the option that best describes your experience.)

   ○ Within the past month
   ○ Within the past year
   ○ Within the past 2 years
   ○ More than 2 years ago
   ○ Never

5. How often do you engage with the reader comments in on-line news web-sites?
   (Please select the option that best describes your experience.)

   ○ At least once a day
   ○ At least once a week

○ At least once a month
○ Very rarely (i.e. more than one month intervals between visits)
○ Never

6. a) When you engage with reader comments, how many individual comments do you typically read?
   ○ 1-5
   ○ 6-10
   ○ 11-20
   ○ 21-50
   ○ More than 50
   ○ Don't know
   ○ Question does not apply to me – I don't engage with reader comments

   b) If you engage with (e.g. read or skim) comments in on-line news, please tell us a bit about why you do so? (*optional*)
   (e.g. "If I am really interested in a new article, I like to see what people have to say in response to the article"; etc.)

   c) If you *do not* engage with (e.g. read or skim) comments in on-line news, as part of your work, please tell us a bit about why you do not? (*optional*)
   (e.g. "I don't have time"; etc.)

7. How often do you post comments?
   (Please select the option that most accurately describes your experience.)
   ○ At least once per day
   ○ At least once per week
   ○ At least once per month
   ○ Very rarely (i.e. more than one month intervals between posts)
   ○ Never

8. If you post comments on a regular basis then approximately how many posts do you typically make in a day:
   ○ 1-5
   ○ 6-10
   ○ 11-20
   ○ More than 20
   ○ Don't know
   ○ Question does not apply to me

9. a) Please tell us a bit about how you post comment as part of your work.
   (Please select **one or more** options from the list below.)
   ☑ I never post comments
   ☑ I often post a comment in response to something I have read in an article
   ☑ I often post a comment in response to something I read in the comments

☑ When posting a comment I typically begin a new thread
☑ When posting a comment I only occasionally begin a new thread
☑ When posting a comment, I typically add the comment to an existing thread
☑ When posting a comment, I only occasionally add the comment to an existing thread
☑ If someone posts a reply to one of my comments, I rarely post a reply back to them
☑ If someone posts a reply to one of my comments, I often post a reply back to them
☑ None of the above

b) If you post comments as part of your work, please tell us a bit about how and why you post comments? (*optional*)
(e.g. "I post comments to correct inaccuracies I see in the article"; "I post comments on topics I am interested in"; etc.)

e) If you *do not* post comments as part of your work, please tell us a bit about why you do not? (*optional*)
(e.g. "I don’t have time"; "I don't feel comfortable putting my views on-line"; "I am worried about 'trolls'", etc.)

Thank you, you have completed the background questions. Please proceed to the main questionnaire.

SENSEI: Use Case Questionnaire

Introduction

In SENSEI (www.sensei-conversation.eu), we are developing new software for helping people make better sense of the reader comments in on-line news.

As part of this work we have imagined several scenarios (also known as "use cases"), which illustrate various ways that our software might help different possible users, such as your self.

In this questionnaire we invite your feedback on these scenarios - your responses will help us to design software that is as useful as it can be.

Before we begin, we would like you to tell us a bit about your experience of using on-line news and comment.

Use Case Questionnaire

In this questionnaire, we present each scenario or "use case" in turn, together with a set of questions.

For each scenario, we identify one or more **user groups**. These include:

1. People who work in the news media, e.g. reporters, sub-editors, etc.
2. Members of the public who engage with news and comments. In this second group we distinguish between news and comment readers (i.e. people who read news and/or comments but very rarely post comments) and comment providers (i.e. people who read news and/or comments and who contribute reader comment on a regular basis).

We also give a **user goal**, i.e. something a user might want to achieve when engaging with news and comment.

We then sketch what the **SENSEI software** might do to help the user achieve their goal.

A set of questions then follows.

We invite you to first carefully read the **Use Case**, then to answer the questions that follow it and finally to add any further comment you may have in the boxes provided.

We greatly value your feedback, so do please comment as freely as possible!

Use Case 1: The Town Hall Meeting Summary

Please indicate how much you agree with the following statements and feel free to comment further in the space provided.

Please respond to the best of your knowledge for all user groups.

User Goal: “To gain an understanding of the key content of a news article and associated set of reader comments”.

1.1 The user goal feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know

Other – please specify other roles in the news media that engage with news and comment and may have this goal:

SENSEI Functionality

1.2 A SENSEI “**THM summary**” would be very useful for the user groups, when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know

Other – please specify other roles in the news media that engage with news and comment and would find this functionality useful:

1.3 Please list as many activities as you can, where **you** in **your current role** as a news-media professional would use a SENSEI “**THM summary**”.

We have given an example as a starter. You may select and add the example to the answer table below if relevant to you. Once in this table you may edit the example as you see fit.

For any additional activity, please indicate your **role, activity**, and also say a bit about ***why*** you would find it useful.

Example:

Role	**Activity**	**Why SENSEI functionality is useful**	
Reporter	Preparing for a meeting with a sub-editor about previous and future story assignments.	A THM summary would provide a quick overview of the readers' comment on and response to recently published stories, to use in discussion with a sub-editor.	***Add to Answer Table***

Activities:

Role	**Activity**	**Why SENSEI functionality is useful**

1.4 Please let us know any further comments, in particular, what you **like** about the use case, what you **dislike**, any things you would **change** and/or anything you think should **add** to the use case:

Use Case 2: Updating/Extending the Coverage of the News Story

Please indicate how much you agree with the following statements and feel free to comment further in the space provided.

Please respond to the best of your knowledge for all user groups.

User Goal: “To gather from the comments any additional information that refers directly to the article, e.g. reports of factual errors, elaboration, recommendations for follow up, etc., with a view to updating the article and extending the coverage of the news story. (Assertions of opinion, off-topic comment, etc. are excluded).”

2.1 The user goal feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know

Other – please specify other roles in the news media that engage with news and comment and may have this goal:

SENSEI Functionality

2.2 A SENSEI “**Update and Extend**”report, which updates and extends a news article via reader contributions, would be very useful for the user groups, when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know

Other – please specify other roles in the news media that engage with news and comment and would find this functionality useful:

2.3 Please list as many activities as you can, where **you** in **your current role** as a news-media professional would use a SENSEI "Update and Extend" report.

We have given an example as a starter. You may select and add the example to the answer table below if relevant to you. Once in this table you may edit the example as you see fit.

For any additional activity, please indicate your **role, activity**, and also say a bit about ***why*** you would find it useful.

Example:

Role	**Activity**	**Why SENSEI functionality is useful**	
Reporter	Would use a SENSEI "Update and Extend" report when preparing a new/updated version of a news article.	This report would save time in checking errors, adding background and perspective; may give ideas for follow-up story.	***Add to Answer Table***

Activities:

Role	**Activity**	**Why SENSEI functionality is useful**

2.4 Please let us know any further comments, in particular, what you **like** about the use case, what you **dislike**, any things you would **change** and/or anything you think should **add** to the use case:

Use Case 3: Backgrounding: Looking in Greater Depth at a Comment Poster

Please indicate how much you agree with the following statements and feel free to comment further in the space provided.

Please respond to the best of your knowledge for all user groups.

User Goal: “To build a picture of a comment provider (aka “poster”) based on other comments he/she has made.”

3. 1 The user goal feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know

Other – please specify other roles in the news media that engage with news and comment and may have this goal:

SENSEI Functionality

3.2 A SENSEI “**Comment Poster Profile**” would be very useful for the user groups, when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don’t know

Other – please specify other roles in the news media that engage with news and comment and would find this functionality useful:

3.3 Please list as many activities as you can, where **you** in your current **role** as a news-media professional would use a SENSEI “**Comment Poster Profile**”.

We have given an example as a starter. You may select and add the example to the answer table below if relevant to you. Once in this table you may edit the example as you see fit.

For any additional activity, please indicate your **role, activity**, and also say a bit about ***why*** you would find it useful.

Example:

Role	**Activity**	**Why SENSEI functionality is useful**	
Reporter	Writing a follow-on story, and reviewing the comments to see if there are any ideas/issues which might be useful to include/pursue in the follow-up.	A "Comment Poster Profile" would help a reporter to find out a bit more about the source (i.e. the comment poster) and to assess its authority/possible bias. He/she would then be more confident in using information provided by the comment in research/content for the new story.	***Add to Answer Table***

Activities:

Role	**Activity**	**Why SENSEI functionality is useful**

3.4 Please let us know any further comments ,in particular, what you **like** about the use case, what you **dislike**, any things you would **change** and/or anything you think should **add** to the use case:

Use Case 4: Finding Similar, Related or Redundant Postings

Please indicate how much you agree with the following statements and feel free to comment further in the space provided.

Please respond to the best of your knowledge for all user groups.

User Goal: “To find other **similar comments** (e.g. comments that make the same point or are closely related in content to a comment of interest).”

4. 1 The user goal feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know

Other – please specify other roles in the news media that engage with news and comment and may have this goal:

SENSEI Functionality

4.2 A SENSEI “**Similar Comment**” report would be very useful for the user groups, when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	☐ Don’t know

Other – please specify other roles in the news media that engage with news and comment and would find this functionality useful:

4.3 Please list as many activities as you can, where **you** in **your current role** as a news-media professional would use a SENSEI “**Similar Comment**” report.

We have given a few examples as a starter. You may select and add these examples to the answer table below if they are relevant to you. Once in this table you may edit the example as you see fit.

For any additional activity, please indicate your **role, activity**, and also say a bit about ***why*** you would find it useful.

.

Example:

Role	**Activity**	**Why SENSEI functionality is useful**	
Comment editor	Seeking to make an "editor's recommendation" in the comments.	A "Similar Comment" report would help a person to establish if the comment was unique, or if others had made the same point or share the opinion; if the comment was representative of many similar comments - is it the "best" example to use from a set of similar comments? He/she could include this information in any recommendation.	***Add to Answer Table***
Reporter	Writing a follow-on story, and reviewing the reader comments on an article to see if there is any information that might be useful.	A "Similar Comment" report would very rapidly assemble similar comments and one could then compare these with the comment of interest. This is very time consuming/difficult to do with current on-line comment systems. This perspective would help a reporter establish if the issue is something widely known, or not; if it is likely to be true; and may suggest whether there were strong feelings on the issue. This would help the reporter decide if information in the comment merits follow-up.	***Add to Answer Table***
Reporter	Looking at what people are saying in response to a recent article he/she had written.	A reporter could rapidly compare any interesting comments with comments in a "Similar Comment" report. This would be very time consuming/difficult to do with current on-line comment systems. The perspective provided by the similar comments would indicate if others share the feeling/ are making the same point as in the comment of interest.	***Add to Answer Table***

Activities:

Role	**Activity**	**Why SENSEI functionality is useful**

4.4 Please let us know any further comments ,in particular, what you **like** about the use case, what you **dislike**, any things you would **change** and/or anything you think should **add** to the use case:

Use Case 5: Identifying Trends in Reader Comments

Please indicate how much you agree with the following statements and feel free to comment further in the space provided.

Please respond to the best of your knowledge for all user groups.

User Goal: “To determine which topic(s) in a specified time period (e.g. week, quarter, year, etc.) have elicited a significant response from the comment posting community.

In particular, to identify news article topics with: very high or low volumes of reader comment; with the most emotive reader content; with the most polarized opinion and to establish what topics emerge in the comments”.

5. 1 The user goal feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments.

Reporter	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Sub-editor	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Comment editor	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Editor	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
News & comment reader	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Comment provider	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know

Other – please specify other roles in the news media that engage with news and comment and may have this goal:

SENSEI Functionality

5.2 A SENSEI “**Trend**” report would be very useful for the user groups, when they engage with on-line news and comments.

Reporter	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Sub-editor	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Comment editor	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Editor	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
News & comment reader	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know
Comment provider	:	Strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	Strongly agree	☐ Don’t know

Other – please specify other roles in the news media that engage with news and comment and would find this functionality useful:

5.3 Please list as many activities as you can, where **you** in **your current role** as a news-media professional would use a SENSEI "**Trend**" report.

We have given an example as a starter. You may select and add the example to the answer table below if relevant to you. Once in this table you may edit the example as you see fit.

For any additional activity, please indicate your **role, activity**, and also say a bit about ***why*** you would find it useful.

.

Example:

Role	**Activity**	**Why SENSEI functionality is useful**	
Comment editor	Writing a report on what the comment community is talking about, based on comments related to an on-going story in the news, e.g. the run up to "The Scottish Referendum".	A "Trend" report will help by showing which issues elicited the most significant response from the comment community and provide a rapid way of assessing and comparing the community feeling and opinion on different issues.	***Add to Answer Table***
Editor	Preparing for a meeting about the paper's future content policy.	A "Trend" report will establish which topics elicited the most significant response from the comment posting community and provide a rapid way of assessing and comparing the community response to different topics in the news. This will help editorial staff to decide which topics to prioritise in future.	***Add to Answer Table***

Activities:

Role	**Activity**	**Why SENSEI functionality is useful**

5.4 Please let us know any further comments ,in particular, what you **like** about the use case, what you **dislike**, any things you would **change** and/or anything you think should **add** to the use case:

Use Case 6: Making Content from the Comments

Please indicate how much you agree with the following statements and feel free to comment further in the space provided.

Please respond to the best of your knowledge for all user groups.

User Goal: "To obtain a view of the 'best' content in the comments, ranked by various criteria".

6. 1 The user goal feels authentic/true to life, i.e. the specified user groups may want to do this when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know

Other – please specify other roles in the news media that engage with news and comment and may have this goal:

SENSEI Functionality

6.2 A SENSEI "**Comment Digest**" would be very useful for the user groups, when they engage with on-line news and comments.

Reporter	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Sub-editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Comment editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Editor	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
News & comment reader	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know
Comment provider	:	Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly agree	□ Don't know

Other – please specify other roles in the news media that engage with news and comment and would find this functionality useful:

6.3 Please list as many activities as you can, where **you** in **your current role** as a news-media professional would use a SENSEI "**Comment Digest**".

We have given a few examples as a starter. You may select and add these examples to the answer table below if they are relevant to you. Once in this table you may edit the example as you see fit.

For any additional activity, please indicate your **role, activity**, and also say a bit about ***why*** you would find it useful.

.

Example:

Role	**Activity**	**Why SENSEI functionality is useful**	
Comment editor	Seeking sets of news articles and comments to summarise in an editorial in, e.g., The Guardian's "comment is free" page or the Saturday Guardian Letter's Page, from a day's/week's worth of comment and news.	The "Comment Digest" would automatically select and summarise the top 5 sets of comments based on a range of factors, saving a significant amount of time and effort. The SENSEI digest would make it easy to compare the individual THM summaries of the articles plus comments. An individual SENSEI THM summary could be the basis for the editorial content - staff could modify or elaborate on the summary if necessary.	***Add to Answer Table***
Comment editor	Selecting "editors picks" from the comments for a day's/week's worth of comment.	The "Comment Digest" provides a list of top ranked comments in an instant. The SENSEI THM summaries of the top ranked sets of comments, may be a good starting point for seeking further interesting comments. They may suggest further comments to consider.	***Add to Answer Table***

Activities:

Role	**Activity**	**Why SENSEI functionality is useful**

6.4 Please let us know any further comments ,in particular, what you **like** about the use case, what you **dislike**, any things you would **change** and/or anything you think should **add** to the use case:

Comparing Use Cases

We would like to know which of the proposed SENSEI functionalities you would most like to have available in on line news and reader comments, based on how useful you believe the functionality would be to you, in your daily work.

Please rank the 6 Use Cases in order of how useful you believe the functionality they describe would be to you in your work:

1. The Town Hall Meeting (THM) Summary
2. Update and Extend Report
3. Comment Poster Profile
4. Similar Comment Report
5. Trend Report
6. Comment Digest

Rank	Use Cases
1 (would be most useful to me)	
2	
3	
4	
5	
6 (would be least useful to me)	

Please comment on your preferences:

Appendix E: Results for Social Media Use Cases

In this appendix, we report the results for the following 6 social media use cases:

- Use Case 1: The Town Hall Meeting (THM) Summary
- Use Case 2: Update and Extend Report
- Use Case 3: Comment Poster Profile
- Use Case 4: Similar Comment Report
- Use Case 5: Trend Report
- Use Case 6: Comment Digest

Table 6 – Number of participants completing the tasks

Tasks	Members of the Public	News-Media Professionals
Use Case 1	18	13
Use Case 2	16	11
Use Case 3	14	8
Use Case 4	14	7
Use Case 5	14	7
Use Case 6	12	7
Comparison of Use Cases	11	7

Table 7 – Demographic of the participating news-media professionals

Role as News-Media Professionals	Number of Participants
Reporter	3
Sub-editor	1
Comment editor	0
Comment moderator	0
Executive/Editorial	3
Educator/Academic	2
Others*	4
TOTAL	**13**

** This category includes participants with the following roles: public relations, private sector worker (communications department), communications officer and social media editor.*

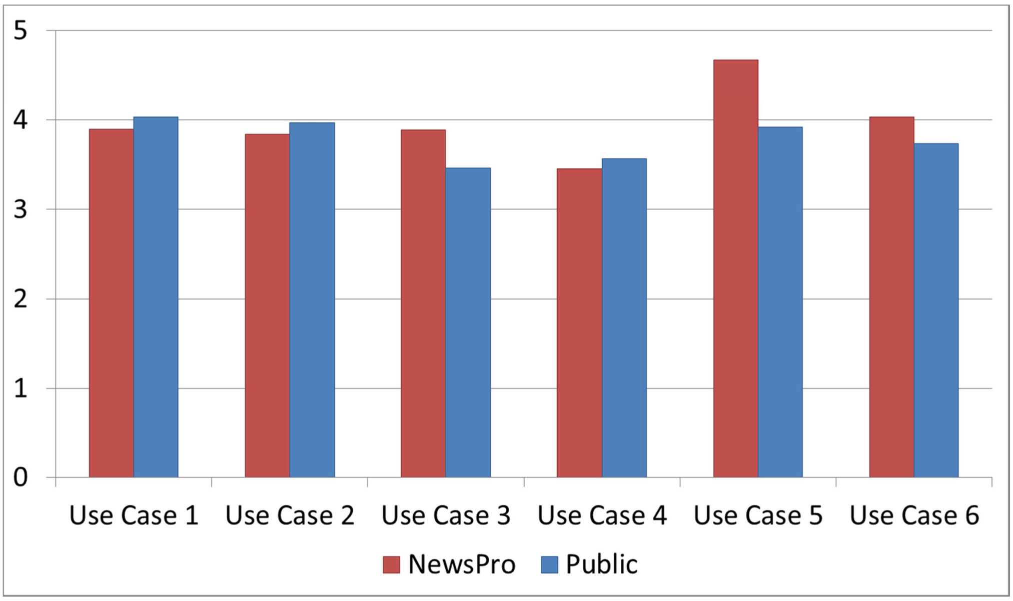


Figure 1 – Question 1 (User Goal)

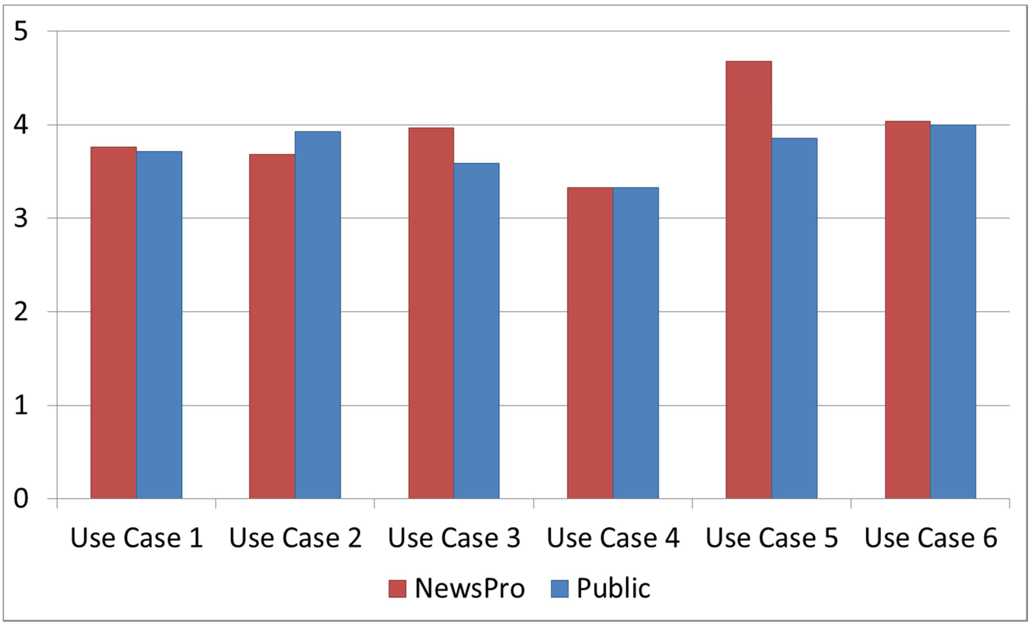


Figure 2 – Question 2 (Functionality)

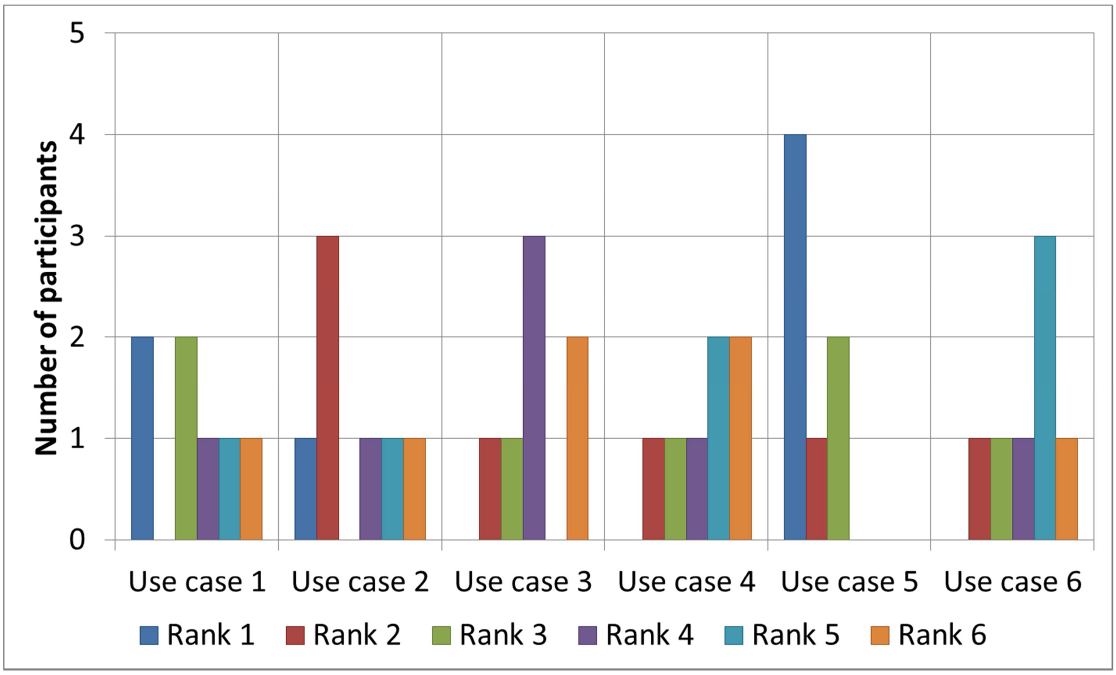


Figure 3 – Rank assignment (newsPro)

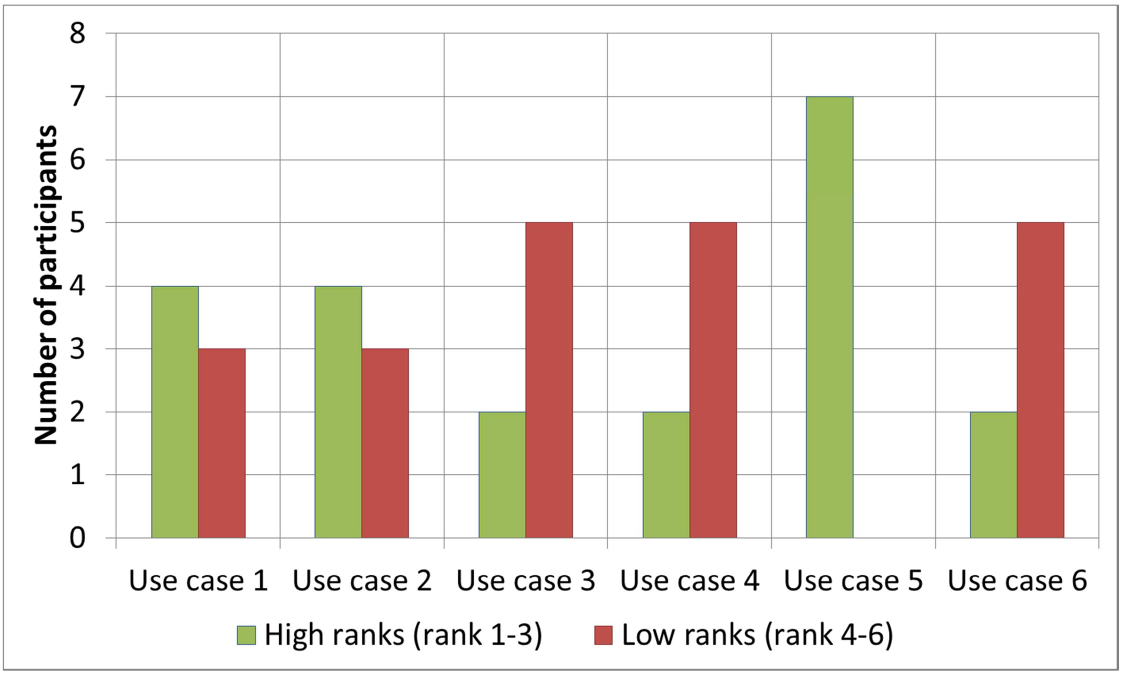


Figure 4 – Rank assignment (newsPro)

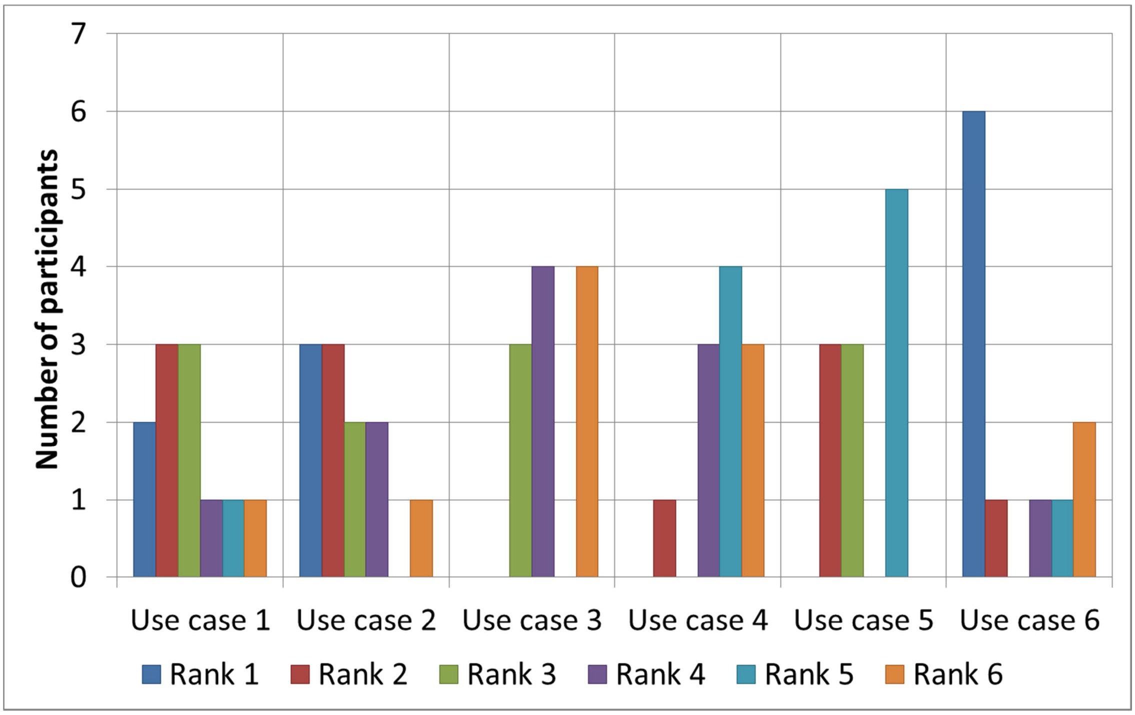


Figure 5 – Rank assignment (public)

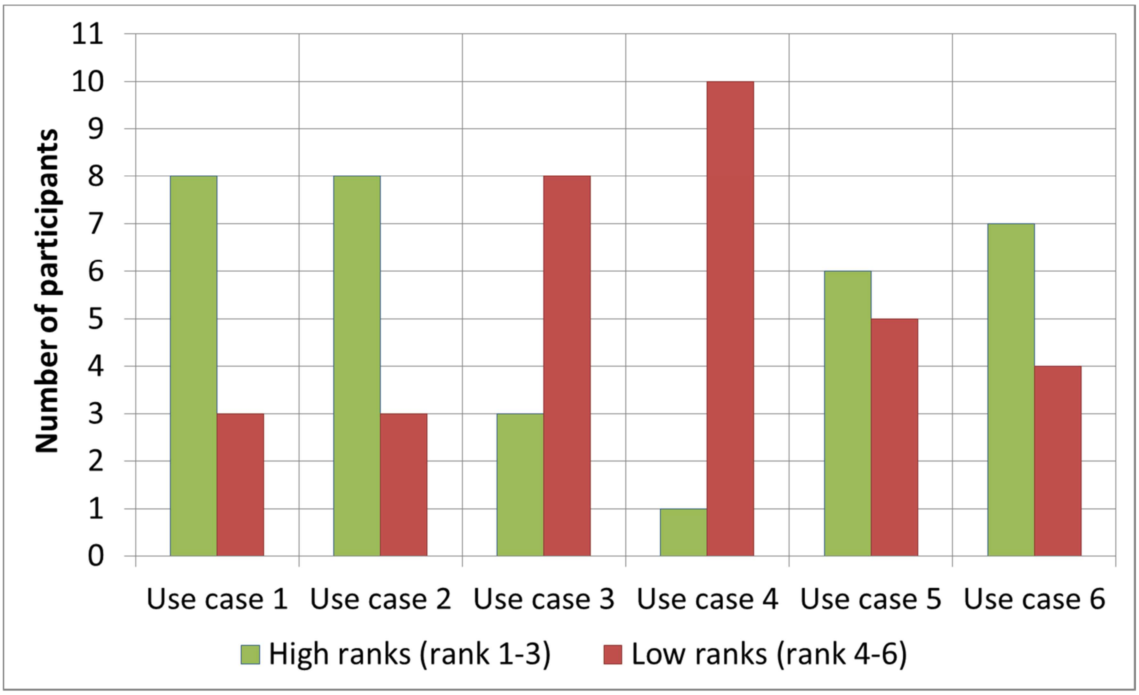


Figure 6 – Rank assignment (public)

Appendix F: Examples of Reader Comment Summarization

This appendix contains examples to illustrate the USFD guided method for writing summaries of reader comments in social media.

Example 1: Comments and Labels

Comment: *Do we still believe that people start smoking because of the packet? All that does is influence people into which brand they smoke, not whether they start smoking or not. I think that they've made the cigarette box into an iconic item today. It's a plain white box ffs, you couldn't make it look cooler if you tried.*

Label: plain packaging for cigarettes; packaging influences brand; plain packaging looks cool;

Comment: *I think the plain packets look very modern and cool actually. And i'm going to invest in a nice new cigarette tin and case. Much more stylish and individual than advertising a brand*

Label: plain packaging looks cool; packaging advertises brand;

Example 2: human authored summary created using our guided method for writing summaries of reader comments

170words

The majority of comments were concerned with whether plain packaging would deter people from buying cigarettes and help to people give up smoking. Opinion was very much divided. Some people argued that plain packaging would help to reduce cigarette sales. Others thought it would not. Some believed the packaging was more important for distinguishing or advertising brands; two thought the minimal white packets made it look "cool". Others suggested that the price of cigarettes would be more important to deter people from smoking. Some commenters pointed to evidence in Australia of decreasing sales where plain packaging had been introduced recently. But there were some who argued that there was evidence to the contrary in Australia. A smaller number of commenters discussed the risk of cancer due to diesel emissions from cars. A few believed that diesel emissions were a more important risk factor for cancer than smoking. One commenter suggested the tobacco lobby was now focussing efforts in developing countries such as Africa. Some others suggested that smoking sales were increasing in these countries.

REFERENCES

Berg, B. L., & Lune, H. (2004). *Qualitative research methods for the social sciences* (Vol. 5). Boston: Pearson.

Cameron, D. (2000). Styling the worker: Gender and the commodification of language in the globalized service economy. Journal of sociolinguistics, 4(3), 323-347.

Clark, C. M., Murfett, U. M., Rogers, P. S., & Ang, S. (2012). Is Empathy Effective for Customer Service? Evidence From Call Centre Interactions. Journal of Business and Technical Communication, 1050651912468887.

Friginal, E. (2008). Linguistic variation in the discourse of outsourced call centres. Discourse Studies, 10(6), 715-736.

Friginal, E. (2009). The language of outsourced call centres: A corpus-based study of cross-cultural interaction (Vol. 34). John Benjamins Publishing.

Ganguly, K. (2009). Managing Customer Hostility in Transnational Call Centres in India. Available at SSRN 1583246.

Grougiou, V., & Wilson, A. (2004). Call centres: the attitudes of the grey market. Journal of Customer Behaviour, 3(2), 147-164.

Grudens-Schuck, N., Allen, B. L., & Larson, K. (2004). Methodology Brief: Focus Group Fundamentals. Extension Community and Economic Development Publications. Book 12. http://lib.dr.iastate.edu/extension_communities_pubs/12.

Jaiswal, A. K. (2008). Customer satisfaction and service quality measurement in Indian call centres. Managing Service Quality, 18(4), 405-416.

Mutton, Andrew and Dras, Mark and Wan, Stephen and Dale, Robert (2007). GLEU: Automatic evaluation of sentence-level fluency. ACL 2007 (7), 344-351.

Nenkova, Ani and Passonneau, Rebecca (2004), Evaluating content selection in summarization: The pyramid method. NAACL-HLT 2004.

Pradhan S, Moschitti A, Xue N, Ng H, Bjorkelund A, Uryupina O, et al 2013 Towards robust linguistic analysis using Ontonotes Proc. Of CONLL, p 147-152

Sjobergh, J. (2007). Older versions of the ROUGEeval summarization evaluation system were easier to fool. IPM 2007, 43 (6) 1500-1505.

Van der Aa, Z,; Bloemer, J.; Henseler, J.. Using customer contact centres as relationship marketing instruments. Service Business, 2013, 1-24.

Vilain, M., Burger, J., Dennis Connolly J.A. & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In Proceedings of the 6th Message Understanding Conference (MUC-6), pp. 45–52. San Mateo, Cal.: Morgan Kaufmann

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association, 67(339), 578-580.