The SENSEI Project: Making Sense of Human Conversations

Giuseppe Riccardi[1(✉)], Frederic Bechet[2], Morena Danieli[1], Benoit Favre[2], Robert Gaizauskas[3], Udo Kruschwitz[4], and Massimo Poesio[4]

[1] University of Trento, Trento, Italy
giuseppe.riccardi@unitn.it
[2] Aix-Marseille University, Marseille, France
[3] University of Sheffield, Sheffield, UK
[4] University of Essex, Colchester, UK
http://www.sensei-conversation.eu

Abstract. Conversational interaction is the most natural and persistent paradigm for personal and business relations. In contact centres customer spoken conversations are handled daily. On social media platforms conversations are delivered in different forms, lengths and for different purposes. In both cases, conversations have little impact on the intended target listeners, due to the volume, velocity and diversity (media, style, social context) of the document streams (spoken conversations and blog posts). Most language analytics technology is limited in that it performs keyword search, which does not provide automatic descriptions of what happened, who said what, which opinions are held on what subject, in a coherent, readable and executable form. In the SENSEI project we plan to go beyond keyword search and sentence-based analysis of conversations. We adapt lightweight and large coverage linguistic models of semantic and discourse resources to learn a layered model of conversations. SENSEI addresses the issue of multidimensional textual, spoken and metadata descriptors in terms of semantic, para-semantic and discourse structures. Automated generation of readable analytics documents (summaries) will support end-users in the context of large data analysis tasks. Summarization technology developed in SENSEI has been evaluated with respect to users' task requirements and performances in the context of contact centre and social media conversations.

Keywords: Summarization · Spoken dialogue · Social media · Language analytics

1 Introduction

Conversational interaction is the most natural and persistent paradigm for personal and business relations. Vast amounts of data of this type are already available to business, yet current language analytics technology only offers limited support. Data analysts facing such a data deluge, need to be able to extract and

J.F. Quesada et al. (Eds.): FETLT 2015, LNAI 9577, pp. 10–33, 2016.
DOI: 10.1007/978-3-319-33500-1_2

summarize relevant information from large quantities of this most fundamental form of human linguistic behaviour. For example, in contact centres millions of spoken conversations are handled daily to provide vital support to business units and their customers. However, a call centre analyst aiming to identify areas for improvement by examining the data collected by her/his company will only be able to study a tiny fraction of such data due to the limitations of speech analytics technology. Similar problems limit the analysis of comment threads on social media platforms, a new type of multiparty conversation in which hundreds of millions of blog posts and related comments are generated both in generalist (e.g. Twitter) or proprietary platforms (e.g. news websites). A journalist wanting to engage with his/her readers by following such threads will be quickly overwhelmed by the amount of data produced. Both types of conversations have limited impact on the intended target listeners due to the volume, velocity and diversity (media, style, social context) of the document streams (spoken conversations and blog posts). The SENSEI vision is to drive forward conversation analytics technology by addressing the state-of-the-art limitations, i.e. to develop analytics technologies that (1) understand conversations at a much deeper level, in particular taking account of para-semantic aspects of conversation (2) automatically generate a range of summary outputs to suit the range of end-users with a stake (e.g. conversation analysts) in making sense of large volumes of conversational data (3) are adaptable to different conversational channels and different user tasks.

This is a project review paper and we are going to refer to available studies and results we have achieved at this time and point to the companion website, [36], where the resources, including data, papers, use case design and reports are made available as they are published.

In the following section we will present the SENSEI vision regarding the modeling of summaries in two use case scenarios: (a) contact centre spoken conversations and (b) social media conversations. In Sect. 3 we review the parsing challenges, objectives and recent novel research work and experiments. In Sect. 4 we propose and motivate the conversation summary types in the context of dyadic spoken conversations and multi-party conversations generated on on-line social media platforms. In Sect. 5 we discuss summary evaluation scenarios for the two use cases.

2 Human Conversations

SENSEI's scientific and technology vision is motivated by both an ecological evaluation and the end-user task requirements. Ecological approaches to system evaluation include both observation of data generated by real industry processes as well as real end-user engagement. This is in contrast to largely unsuccessful top-down approaches that push niche and/or early-development technology into the development pipeline. To this end SENSEI has identified two use cases that are prime exemplars of the diverse space of applications for conversation analytics in the consolidated telephony and social media platforms. The two use

cases pose similar technological challenges in terms of language understanding technology in real-world contexts. However such conversations occur over significantly different media (speech vs text) and social context (dyadic real-time conversations vs n-adic non-real time conversations). For each use case we have defined summary categories that we will propose and discuss in Sect. 4. Such summary categories may cover existing document types as well as new types that will address limitations of current analytics technology. Last but not least, the two use cases will allow us to instantiate both multimedia and cross-media investigation and technology development.

Call Centre Use Case. In outsourced call centres, large corporations outsource their customer touch-point to a hosting call centre. The in-coming and outgoing calls may be monitored in real time, or recorded for later review. The monitoring is done by human evaluators for small random call samples (much less than 1 %). Their job is to track indicators of call quality and agent efficiency. The call centre's corporate client may require reporting in different aggregated forms according to, e.g., the topic of the calls or, in other words, what their customers are asking about, or the emotional content of the call, e.g. concerned or frustrated user. The services provided by the human analysts and evaluators are very expensive in some cases or not feasible in others because of the data deluge or task complexity. The end-users of SENSEI analytics results are professional analysts working in call centres. Depending on the target of their evaluation (e.g. monitoring of agent efficiency, control of call quality, identification of call topic, evaluation of user satisfaction, evaluation of agent training needs), they will be able to profit from the different categories of summaries and reports generated by SENSEI systems.

Social Media Use Case. In a news publisher website such as *The Guardian* or *Le Monde*, journalists publish articles on different topics from politics and civil rights to health, sports and celebrity news. The website design supports the publication and consumption of original news articles and at the same time facilitates user-involvement via reader comments. Increasingly, in a period of disruptive change for the traditional media, newspapers see their future as lying in such conversations with and between readers, and new technologies to support these conversations will become essential. In this scenario there are a number of potential users:

- news readers and the originating journalist want to gain a structured overview of the mass of comments, both in terms of the sub-topics they address and their connection with the original article and in terms of the opinions (polarity and strength) the commenters hold about these topics;
- news readers who join a forum discussion need to be empowered so that they can respond to the originating article and/or to a sub-set of earlier comments that may be relevant to their own personal view on the matter;
- editors or media analysts may need a more widely scoping analysis.

At present none of these users can effectively exploit the mass of comment data – frequently hundreds of comments per article – as there are no tools to support

them in doing so. What they need is new tools to help them make sense of this data deluge. In this scenario, therefore, SENSEI end-users will be news comment readers, news comment authors, journalists and editors/media analysts. Users in these categories will benefit from the various types of summaries and reports generated by SENSEI systems.

Figure 1 shows the overall architecture of SENSEI workflow in the context of the two use case scenarios. SENSEI conversational data are taken from call centres and social media platforms. They are parsed and annotated with semantic, para-semantic and discourse level descriptors and aggregated to yield summaries for end-users in the form of conversational-oriented summaries (e.g. topics T_i categorized using domain ontologies or multimedia extractive summaries), blogger-oriented summaries describing groups (e.g. group, G_i, orientations towards topic T_i), user-defined ad hoc reports (e.g. composition of semantic and para-semantic aspects) and rated questionnaires (e.g. call quality monitoring forms).

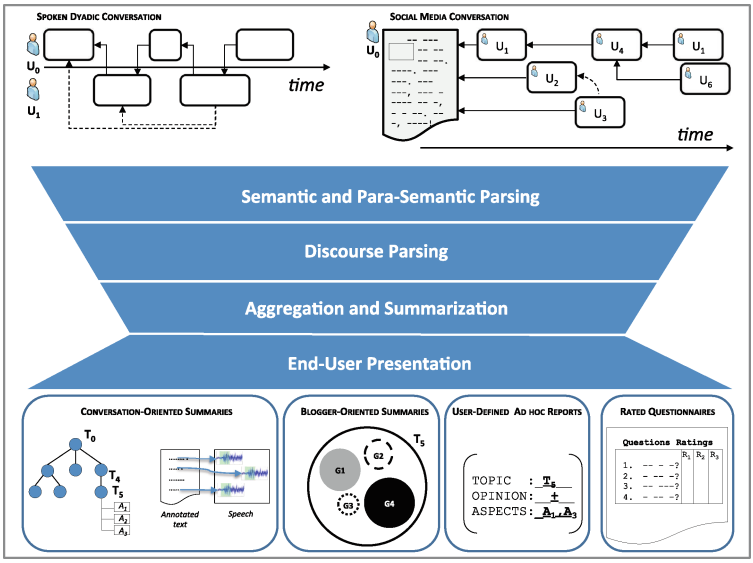

Fig. 1. SENSEI conversational analysis, parsing and summarization work-flow. Conversations are automatically annotated with semantic, para-semantic, discourse level descriptors and aggregated to yield summaries for end-users. The summaries are in the form of conversational-oriented summaries (e.g. topics T_i categorized using domain ontologies or multimedia extractive summaries), blogger-oriented summaries describing groups (e.g. group (G_i) orientations towards topic T_i), user-defined ad hoc reports (e.g. composition of semantic and para-semantic aspects) and rated questionnaires (e.g. call quality monitoring forms)

3 Parsing Human Conversations

3.1 Semantic Parsing

Semantic parsing is the process of producing semantic interpretations from words and other linguistic events that are automatically detected in a text conversation or a speech signal. Many semantic models have been proposed, ranging from formal models encoding deep semantic structures to shallow ones considering only the main topic of a document and the concepts or entities occurring in it. For Open Domain Semantic Parsing, generic purpose semantic models can be used, such as FrameNet or Abstract Meaning Representation (AMR). Once this generic meaning representation is obtained, a translation process trained on a small annotated corpus can be applied for projecting generic predicates and concepts to application specific ones. This kind of approach can help to reduce the need for large application-specific annotated corpora for training Natural Language Understanding (NLU) models by taking advantage of generic resources already available. This is the approach followed in SENSEI.

Deep Neural Network Models. Recent computational representations based on a continuous vector space for words have been used to overcome the need for annotated corpora by taking advantage of very large collections of unlabeled data to model both semantic and syntactic information. In particular researchers in Natural Language Processing have focused on learning a dense low dimensional (hundreds) representation space of words [38,47,53], called embeddings. The benefits of such representations are (1) that they offer a lower computational complexity when used as input of classifiers such as neural networks, and (2) that words with similar properties have similar representations, allowing for better generalization from subsequent models, e.g. for words not covered by targeted task training data. This strategy has been applied successfully for many classical NLP tasks such as information retrieval, language modeling, machine translation, as part-of-speech tagging, named entity recognition, syntactic parsing, semantic role labeling, etc.

Three main characteristics make DNN-based models good candidates for building NLU models:

- the use of a large amount of unlabeled data for learning word representations when dealing with a limited amount of in-domain data [58];
- the joint optimization of DNN over several NLP tasks;
- the ability of Recurrent Neural Networks (RNN) to maintain contextual information through sequence decoding with a memory model such as the Long Short Term Memory model [55].

This last characteristic is particularly relevant to SENSEI as one of its main foci is on the representation of conversational context in semantic parsing models.

One drawback of embeddings and DNN for semantic parsing on conversational data, as noted by [37], is the fact that they are usually obtained on very large written text corpora covering generic domains, such as news articles or

Wikipedia pages, although semantic parsing systems are dealing with spontaneous speech and non-canonical text on specific domains. To overcome this limitation in SENSEI we have proposed several adaptation methods along three dimensions: cross-media, cross-domain, cross-language.

These adaptation methods in SENSEI follow a common strategy:

- Open domain Semantic Parsing with generic semantic models such as FrameNet.
- Joint use of a large amount of unlabeled data as well as rich linguistic resources in word embedding representation approaches when dealing with no or little in-domain annotated data.
- Adaptation to a new media/domain/language in the embedding space thanks to little adaptation data.

For example in [56] we address the cross-media/cross-domain issues by both adapting an embedding space trained on Wikipedia thanks to a small adaptation corpus containing spoken transcriptions corresponding to the call-centre we were dealing with; then by generalizing this adaptation to all words of the original embedding space, in particular to those not occurring in the adaptation corpus. A comparison of CRF and Neural Network methods is given in Fig. 2 for the semantic frame tagging task on the SENSEI call-centre corpus. This figure presents the results obtained by increasing the amount of adaptation data. CRF and NN only use word features. CRF++ uses as well Part-Of-Speech features; NN+ correspond to the adaptation process proposed in SENSEI.

3.2 Para-Semantic Parsing

Para-semantic parsing aims at analyzing paralinguistic features of human conversations and complements the semantic analysis of a conversation. Such features include turn-taking descriptors (e.g. speech overlap), speech rate, speech quality and pitch segmental statistics for spoken data and non-verbal cues such as text format features and emoticons for social media data. In SENSEI our goal is to investigate the relation with semantic features and aim at a joint or composite model. In social media analysis, most of the previous work on para-semantic traits has been done in the framework of Sentiment Classification, further divided into Opinion Detection and Sentiment Polarity Classification. An opinion can be defined as a quadruple: author (opinion holder), target audience, an object of the opinion, and semantic orientation (polarity) of the opinion (optionally also intensity of sentiment). The main focus of sentiment analysis research has been user reviews, to a much lesser degree blogs and forums, and significantly less dyadic or multiparty conversations. Thus, the analysis is generally limited to identification of semantic orientation, where supervised machine learning with bag-of-words models yields satisfactory performance. In the analysis of conversations, opinion holders, target audience and objects of opinion play a crucial role. Notable exceptions in the field are works that do stance classification in online debates or dialogues [Somasundaran and Wiebe, 2009, 2010]; they show

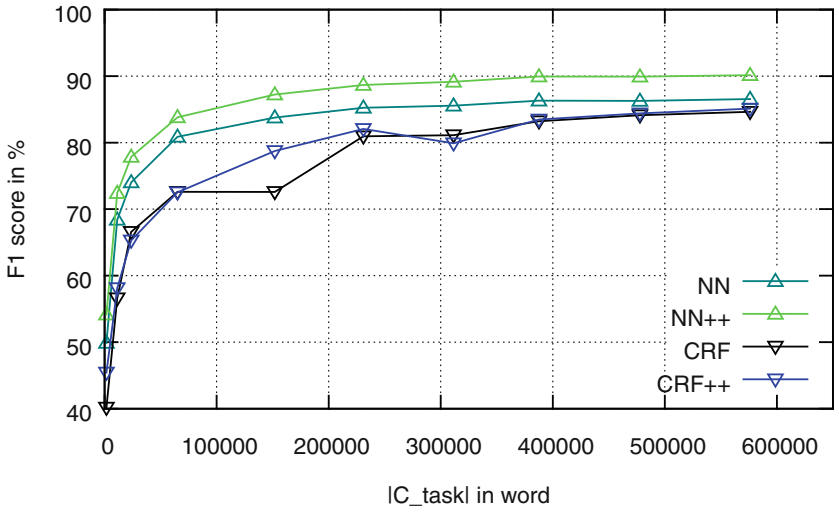

Fig. 2. Semantic Frame tagging performance (F-score) as function of the increasing amount of adaptation data. Comparison of CRF and baseline Neural Network approaches are shown as well the adaptation process proposed in SENSEI, denoted NN++.

that sentiment analysis of conversations requires a richer set of features, such as dialogue acts and discourse-based features. However, even in these works the full potential of discourse analysis is not explored, e.g. only discourse connectives are considered.

In spoken conversations in the last twenty years there has been a growing interest in and research work on *affective computing*, a comprehensive term including research on computational models of emotions, affect, personality and attitudes. However the analysis of emotions and computational models of them has been done in isolation from the semantic or discourse descriptions of human conversation. Last but not least, the emotion space (e.g. Ekman categories) is limiting for the richness and diversity of human conversations observed in-the-wild, such as public forums, business and personal communications.

Affective Scenes. In SENSEI, we have introduced the concept of affective scenes [42]. An affective scene is an emotional episode where one individual is affected by an emotion-arousing process that (a) generates a variation in their emotional state, and (b) triggers a behavioral and linguistic response. The concept of affective scenes has been proposed to explain the unfolding of basic emotions in conversations and applied to operator-customer call analysis. In Fig. 3 we show a state representation of the affective scene. Starting from an initial state (e.g. customer-operator greeting) one of the two speaker may manifest first his/her emotion (e.g. frustation) followed by transition into other states (e.g. anger). The conversation will end into either a *positive* or *negative* state.

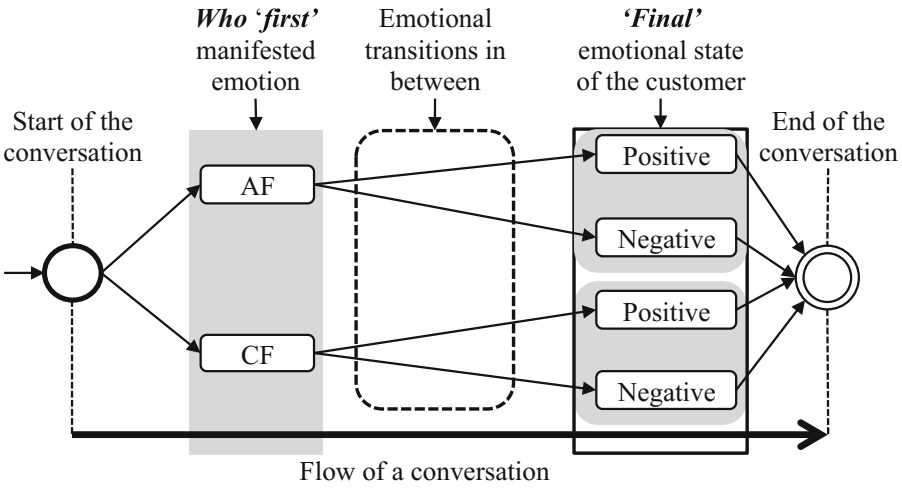

Fig. 3. State traces of affective scenes.

Affective scenes complement the linguistic scenes and their descriptions are being integrated to give a rich and complete description of the human conversations. An interesting extension of the two-party affective scene may be explored for multi-party conversations occurring in social media platforms.

Speech Overlaps. Another relevant topic we have investigated is *speech overlaps* and their semantic and discourse function. Speech overlaps are important events in spontaneous spoken conversations. In contact centers, speech overlap segments account for less than 10 % of the spoken segments [39] and they are required for *stitching* together the speech acts of speakers. Overlapping speech may reflect many aspects of discourse dynamics as well as emotional states. In [39,40] we have focused on the pragmatic role of *competitive* or *non-competitive* overlaps and the roles of speakers in the act of overlapping. Further research will include the investigation of speech overlaps with respect to the semantic as well as affective description of human dialogues.

3.3 Discourse Structure and Coreference

It has long been known that the structure imposed on discourse by the relations underpinning its (relational) coherence is key to the human ability to recall and summarize information (e.g., [16,28]). This link was a key motivation for the early work on discourse structure and discourse parsing [13,14]. More recently, it has been shown that the information about entity coherence information [9,19] that can be extracted from text by intra-document and inter-document coreference resolution algorithms [20] also helps single-document and multi-document summarization [24,25] by identifying the main entities of a document or a collection of documents. These findings made the analysis of discourse structure and

coreference a key aspect of SENSEI. In the following subsections we describe the main research topics we have addressed.

Domain Adaptation for Discourse Parsing. Much of the early work in discourse parsing was based on Rhetorical Structure Theory [13,14], but much of the modern work in the area has been spurred by the creation of the Penn Discourse Treebank (PDTB) [21], based on a connective-driven theory of discourse structure. The PDTB, however, consists mostly of text, and therefore there has been limited work on applying discourse parsing to spoken conversations, and even less to social media. Discourse structure in conversations differs in a number of respects from that of text. For instance, dialogue has more pragmatically motivated relation types when compared to written text, such as Interruption (speaker couldn't complete an utterance). Work on discourse parsing in SENSEI has therefore focused on adapting methods developed for discourse parsing to take into account the nature of speech [26,41].

Argument Structure. Among the relations found in conversations, those that specify the structure of arguments were expected to be of particular interest to SENSEI. Social media such as blogs or commentaries to newspaper articles have an inherently argumentative structure: people agree or disagree with a particular point being made. In order to properly understand such interactions it is essential to recognize which of the comments support the point of the commenter and which ones instead are opposed. Argumentation mining has gained increased interest in recent years [18,23,33]; much of this work has been applied to the classification of argumentative propositions in online user comments [1,4].

In SENSEI, we early on identified argument structure as an aspect of discourse structure of particular relevance to the task of summarizing online conversations, and have devoted substantial effort to it, by organizing a shared task on Online Forums Summarization at MULTILING-2015 that has focused on argument structure summarization [5,8] and by creating resources to support the task [2,8]. The shared-task annotation data may be obtained by contacting the consortium at [36].

Coreference. Intra-document coreference is the task of identifying the mentions that refer to the same entity within a document. Annotated corpora for this task became available in the mid-90s, enabling a great deal of research [20]. Recent corpora such as OntoNotes and ARRAU also moved away from annotation schemes motivated entirely by information extraction applications; systems trained on such corpora have been shown to work better for applications such as summarization that rely on some measure of text cohesion [25]. There has only been, however, limited work on coreference in spoken conversations and social media analysis, because of the lack of resources – to our knowledge, prior to SENSEI the LiveMemories-Blog corpus of Italian [22] was the only collection of social media data annotated for intra-document coreference, and we are aware of only one study of intra-document coreference for social media [12]. As in the case of discourse parsing, our primary objective was to develop methods

to adapt models for coreference resolution trained on news to the conversation domain. We have carried out two lines of research in our work in this area. On the one hand, we have created annotated resources to study coreference in online forums, annotating for coreference the English and Italian datasets created for the Online Forums summarization task. On the other end, we have carried out research on domain adaptation for spoken conversations and social media data using our own BART platform, already tested in the 2011 and 2012 CONLL shared tasks [29,30]. Work so far includes adapting BART to run on French conversations [7] as well as work on domain adaptation for social media.

4 Summarization

There is a large body of work on text summarization, but very little that is specifically relevant to the analysis of human conversations in such diverse contexts as speech and social media. Good general overviews of automatic text summarization can be found in [59,60]. In the rest of this section we briefly review the related literature, discuss the novel research problems addressed by SENSEI and present preliminary results.

4.1 Speech Summarization

First approaches to spoken conversation summarization [32,35] have mostly focused on extractive summarization, which consists in selecting relevant utterances from the recordings and displaying their transcript to the user. Those approaches and all the extractive approaches proposed after them [3,6,10,34,59] have shown limitations in that they decontextualize the participants' discourse, and are unable to generalize and relate events discussed over a long time span.

In the call-center domain, in [61] they aim at automatically completing post-call logs, a type of summary generally manually created. The approach consists in filling templates with structured parts (detected from speech recordings) and unstructured parts created with extractive summarization methods. The authors show that call handling times are reduced without compromising log quality. [62] also address the problem of generating call-centre dialogue summaries, but with an unsupervised approach that performs topic induction and extracts utterances under an HMM model. Evaluation is only performed on synthetic dialogues. [63] adopt a different approach which leverages existing pairs of (speech recording, call log) through a method which associates utterances and log words. The method has a negative impact on call log quality, even though it outperforms other automatic baselines.

In SENSEI, we aim at going beyond extractive summarization in order to create abstractive descriptions of the content of conversations. In particular, abstractive summaries of call-centre conversations should be able to yield insight into why the customer called, what was her query, how did the agent solve that problem, was the behavior of the agent appropriate during the dialogue. We call such summaries *synopses*. They have two roles in the project: showing that we

Table 1. Example of synopses written by annotators for a single conversation from the Decoda French corpus [64] which includes calls from citizens enquiryng about public transportation.

Annotator	Synopsis
1	Request for itinerary from suburbs to downtown Paris. The caller wants to understand the fare given by one of his employees.
2	Request for information about the zones to take for a Navigo card for one person living in Chailly-en-Brie to travel in Paris. Zones 1 to 6.
3	An employer is calling the customer service cause he is not very sure about the ticket he has to pay for his employee. His employee is asking him for a sum which doesn't correspond to the fares and so he has the feeling that he is being ripped off.

have reached a sufficient understanding of the conversations, and creating a short textual representation of conversations that can be used to browse call-centre large databases, compare and group similar conversations, and help supervisors find conversations requiring more investigation. Examples of synopses are given in Table 1.

Unlike news summarization, which focuses on locating facts in text written by journalists and selecting the most relevant facts, conversation synopses require an extra level of analysis in order to achieve abstraction. Turn taking from the speakers has to be converted to generic expression of their needs, beliefs and actions. Even though extractive systems might give a glimpse of the dialogues, only abstraction can yield the story of what happens in the conversations.

Recent work on abstractive speech summarization includes modeling text generation as a Markov Decision Process [17] and generating a summary word by word, given a set of sentence clusters from the input. It is reminiscent of the recent trend towards conditioned language models [27,31] which use Recurrent Neural Networks for producing words. A similar approach [15] finds sentence communities through textual entailment and merges them. While those approaches are adequate when large quantities of annotated data are available, they are unsuitable for call-centre conversations which are focused and non-redundant.

Preliminary work on the project has yielded an approach for creating abstractive summaries from conversation transcripts. It uses domain knowledge to fill hand-written templates from entities detected in the conversation transcript using topic-dependent rules. For example, for the public transportation domain, we first cluster conversations by topic, and then write a template for each topic. Each template is a regular language with optional and repeatable parts. Slots are expressed as cross-template variables which need to be filled from the conversation (Table 2).

We performed evaluation on a subset of templates on the CCCS Shared Task for the Decoda corpus [64] using the ROUGE-2 evaluation metric [69]. The abstractive summarization systems are compared to extractive and

Table 2. Example of templates manually created for the Decoda French corpus (translated from French) [64]. We use the regular-expression formalism for denoting optional an repeatable parts.

Topic	Template
Itinerary	Query for itinerary (using $TRANSPORT)? from $FROM to $TO (without using $NOT_TRANSPORT)?. (Take the $LINE towards $TOWARDS from $START_STOP to $END_STOP.)*. Query for location $LOCATION.
Navigo pass	Query for (justification|refund|fares|receipt) for $CARD_TYPE. Customer has to go to offices at $ADDRESS.
Lost&found	$ITEM lost in $TRANSPORT (at $LOCATION)? (around $TIME)?. (Found, to be retrieved from $RETRIEVE_LOCATION |Not found).

abstractive baselines. The extractive baselines are the longest turn of the conversation, the longest turn in the first quarter of the conversation and Maximal Marginal Relevance (MMR). The first abstractive baseline consists of replacing the slot values with a bogus token which is not matched by Rouge during evaluation in order to simulate the worst slot filling system. The second baseline is based on the assumption that named entities play an important role in synopses: it consists in concatenating conversation named entities until the length constraint, without repetition. This baseline achieves a very bad readability, as expected. The topline consists in replacing the slot values with those manually annotated in the reference synopses. Results are summarized in Table 3.

In addition to hand-written templates, which fit well-structured conversations, we have addressed unexpected events through template generation. Following [65], additional templates are learned by extracting frequent patterns from hand-written synopses, generalizing slot variables and filling the templates with entities extracted from the conversation transcript. The generalization and template generation process includes *(a)* aligning synopses to conversation

Table 3. Rouge-2 results of the Decoda synopsis generation systems on a subset of the CCCS test set [64].

System	Rouge-2
Longest turn extract	0.04030
Longest turn @ 25 %	0.04594
MMR extract	0.04490
Hand-written templates + Bogus slots	0.02228
Named entities concatenation	0.09337
Hand-written templates + auto slots	0.10084
Abstractive topline	0.18067

sentences sharing the same semantic frames *(b)* mapping word tokens into their WordNet synsets and *(c)* clustering the generalized synopses to form the final templates.

4.2 Social Media Summarization

Previous work on summarization of text-based conversations and specifically of reader comment in on-line news is even more limited than that on summarization of spoken conversation. Summarization of email threads [54] and chat/on-line discussions [57] are similar tasks but there are critical differences. In the case of reader comment there is an initial news article that readers comment on and the relation of comments to this text is central – there is no direct analogue to this in the case of email or on-line discussion. Furthermore, email and on-line chat tend to involve longer exchanges between smaller numbers of correspondents in a more conventional dialogue form.

A small number of authors have directly addressed the task of summarizing on-line conversations commenting on videos or news articles. Khabiri et al. [50] addressed the task of summarising comments relating to Youtube videos. Ma et al. [51] addressed the task of summarising reader comments in on-line news, specifically *Yahoo! News* with a view to generating "an easy overview of all topics discussed in the comments". Llewellyn et al. [52] address the task of summarising reader comments in *The Guardian* newspaper and follow a similar approach to [50,51], again adopting a three stage process of topical clustering, ranking comments within clusters and then selecting top ranked comments across multiple clusters.

By contrast with earlier work that does not examine what form summaries of reader comments should take, in SENSEI we began by working with end users – journalists, news editors and readers and posters of reader comments – in a comprehensive study to identify use cases surrounding access to information in reader comments [48]. Six use cases were identified, including issue-oriented summaries of a single article+comment set, "blogger-oriented" summaries of all the postings of a single commenter and trend analysis summaries tracking issues across multiple article+comment sets over time.

We have chosen to focus initially on the use case of generating issue-oriented summaries of the comment set associated with a single article, a task bearing similarities to that of a journalist covering a town hall meeting. To support this work we have generated a set of gold standard human-authored summaries for a set of 18 article+comment sets, taking just the first 100 comments for each article [45]. This is the first set of such human-authored summaries for reader comments and the method and tools developed to create it as well as the resulting resource is a significant outcome of the project. Summary authors were given guidelines that, put briefly, instructed them to identify key issues discussed in the reader comments, positions taken with respect to these issues and the emotional tone of the discussion and to aggregate over these when writing their summaries. I.e., summaries were of the form "Many commenters discussed X with most taking stance S while a few took stance T. Other commenters debated Y in a very heated

exchange with ...", capturing the issues discussed, the distribution of views on these issues and the affective character of the discussion. As a side of effect of the summary writing process, summary authors also grouped comments (around the issues discussed) with bi-directional links between comments, comment groups and summary sentences.

In SENSEI we have developed two approaches to automatically generating summaries of single article+comment sets. The first is an *extractive approach* that follows the same general line as previous work: clustering comments by topic, then ranking comments within clusters and finally selecting comments from within clusters to produce a final summary. However, there are several significant differences. First, we have developed a technique to link sentences in comments to sentences in the original article to which they are most similar, or none if the similarity is below a threshold [66]. This capability is used both in clustering (two comment sentences that link to the same article sentence are likely to be in the same cluster) and in summarization, where we have experimented with building summaries from comment clusters in different ways depending on whether or not the cluster contains any comments linked to the article (one might conjecture that summaries linked to the article are more on-topic/serious and hence more likely to contribute to issue-based summaries). Secondly, we have used a different method for clustering, the graph-based Markov Clustering Algorithm [67], leading to clustering results the significantly perform the state-of-the-art LDA-base approaches adopted to date. Finally, we have experimented with many different ranking methods.

Ranking and Extractive Summarization Results. Given a set of comment clusters, extractive summaries may be generated from them in a many different ways. Essentially this comes down to two separate ranking tasks: ranking clusters and ranking sentences within clusters. Summaries are then generated by visiting each cluster in ranked order and selecting from each the top-ranked sentence, until the summary length constraint is reached. We explored three classes of approach.

1. *Baseline Approaches:* No language processing is carried out. Threads are taken to be topically coherent comment groupings, so no clustering is used. Three variations of thread (cluster) sorting were considered: by time of first comment, by number of distinct participants and by number of comments. Comments within threads are sorted by time of posting. In this set of approaches sorting by number of comments worked best (**ParticipantCount-CentroidClosest**).
2. *Basic Text Processing Approaches:* Here again we take threads to be topically coherent comment groupings but consider 5 ways of ranking threads and 2 ways ranking comments within threads. Three ways of ranking threads are the same as used in the baseline approaches and in addition we consider ranking threads by cosine similarity of the thread centroid to the original news article (computed using a standard vector space model with each comment modelled as a vector) and by similarity of the thread centroid to the lead of the news article (first 5 sentences of the article). Within threads comments

Table 4. Summary evaluation results.

System	R1	R2	R-SU4
Human-Human	0.41	0.07	0.13
Time-CentroidClosest-Comment-in-Thread	0.35	0.04	0.10
ArticleLead-Sim-CentroidClosest-Comment-in-Thread	0.42	0.05	0.13
Linked-Cluster-ArticleLeadSim-Summary	0.40	0.04	0.12

are sorted either by time of posting or by cosine similarity of comments to thread centroid. Three of these 10 possible approaches are the same as the baseline approaches. Of the 7 new approaches the one that works best is ranking threads by similarity to the article lead and comments within a thread by similarity to the thread centroid (**ArticleLead-Sim-CentroidClosest-Comment-in-Thread**).

3. *Clustering and Article-Linking Approaches:* The final set of approaches make use of comment-article linking and comment clustering, as described above. A comment cluster is said to link to the original article if any of the comments in it link to the original article. This gives rise to three sets of clusters: linked clusters (all clusters are linked), unlinked clusters (no cluster is linked) and all clusters (linked or unlinked). We experimented with generating summaries from comments taken only from these different cluster sets and found best results were obtained by using just clusters from the linked set of clusters, sorting these cluster by cosine similarity of cluster centroid to article lead and then sorting sentences by cosine similarity to cluster centroid (**Linked-Cluster-ArticleLeadSim-Summary**).

To assess the quality of extractive summarization we use the gold standard summaries described above. We compared the automatically generated summaries against the model summaries using ROUGE [69] and using the standard measures of ROUGE 1 (R1), ROUGE 2 (R2) and ROUGE SU4 (RSU4). ROUGE 1 and 2 give recall scores for uni-gram and bi-gram overlap respectively between the automatically generated summaries and the reference ones. ROUGE SU4 allows bi-grams to be composed of non-contiguous words, with a maximum of four words between the bi-grams. The results of the summary evaluation are shown in Table 4 for the best of class system variants; full details may be found in [46].

The results show that one of the basic text processing approaches works best, one that does not bother with topical clustering but simply takes threads as topic clusters. Two caveats should be made, however. The first is that numerical differences here are small and may not be significant. The second is that as the gold standard summaries are abstractive summaries that feature aggregation over comments, ROUGE, which is fundamentally a lexical overlap measure, may not be appropriate as an intrinsic evaluation measure for this type of summary. The low human-human scores, as compared with the basic text processing approach, may support this sceptical view.

The second approach to summarization of reader comments associated with a single article is a *template-based approach*. Building on the definition of summary type for the issue-oriented or town hall summaries (see [46,48]), we defined a summary template consisting of the article title, a list of main topics discussed in the article and comments, the moods associated with the main topics, an indication of where opinion was consensual or divided, the most central topic and the key contributor to the discussion. The template is filled with data from three different modules: topic extraction, mood prediction and agreement/disagreement detection. Topic model is computed via the hierarchical Latent Dirichlet Allocation (LDA) over each news article and its user comments. The agreement/disagreement detection is based on the relation defined in the CorEA corpus of Italian reader comments *Corriere* [2]. Following the automatic topic linking, mood and agreement-disagreement relations prediction, a final template filling module writes out the template. Individual components on this approach have been evaluated in [46]. The running prototype can be viewed at [36].

On-going work on summarisation in SENSEI is now looking at moving beyond extractive and template-based approaches towards abstractive approaches that will take advantage of work on semantic parsing, parasemantic analysis and discourse and coreference analysis to generate summaries more akin to those that users have specified and that our gold standard exemplifies.

5 Evaluation of Summarization End-User Systems

In Sect. 4.2 above we have discussed the gold standard summary resource we created for evaluating reader comment summaries. This sort of resource is useful for *intrinsic* evaluation of summaries: it allows system developers to assess how close the summaries their systems produce are to what we believe a model summary to be. However, it does not tell us whether our summaries are helpful to end users in some task context. To do the latter we need to specify an *extrinsic* evaluation: a user task, a system or systems to assist the user with the task and metrics for assessing how well a user has performed at the task using the system(s). In SENSEI, the common approach to the extrinsic evaluation is to have the quality and usefulness of the summary to be assessed by the end-users. In the following sections we report on the evaluation frameworks for the speech and social media use cases.

5.1 Speech Use Case

For the evaluation of the SENSEI speech summarization prototype we follow an incremental evaluation model that includes the specification of the tasks, the selection and annotation of exemplar data and the comparative analysis of performances. The process is repeated over the development process of the prototype. Feedback from the evaluation cycles allows the assessment of the performance of the prototype, and the validation of the use cases.

In the speech scenario we have identified the Quality Assurance (QA) supervisors of a call centre as potential end-users. In contact centres the QA supervisors listen to the call and evaluate agents' compliance with the company protocol during the conversations with their customers. Agents' behaviour contributes to the overall quality of the calls, and the QA supervisors score the quality against established contact handling criteria, summarised into a QA monitoring form. In state-of-the-art business processes, the conversations are scored manually and results are recorded in the so-called Agent Conversation Observation Form (ACOF henceforth) [43]. This process may be both time consuming and sometimes inefficient due to the limited amount of calls that QA professionals can listen to every day. One of the goals of SENSEI is to automatically review and score operator-customer calls, and to summarise the features of the agents' behaviour in each call by an automatically generated QA form (e.g. the ACOF). Additionally, as discussed in Sect. 4.1, the goal is the automatic generation of short summaries (synopses) of each call. The speech use case evaluation has been carried over those two tasks.

For the ACOF generation task, the SENSEI prototype classifies the conversations on the basis of aspects of the agent's behaviour, such as the agent's ability to solve the customer problem, their empathic attitude, call resolution effectiveness, and so on. The goal is to evaluate the predictive performance of the SENSEI system in classifying the calls according to the ACOF criteria. We have designed an evaluation task where the automatic ratings assigned by the SENSEI prototype are compared with those assigned by human evaluators. In our case the human evaluators are QA analysts and supervisors. On average, evaluators find the SENSEI prototype is sufficiently accurate for the French and Italian corpus. The Likert ranking for both the Italian and the French corpus was 2.8. Details of this evaluation task can be found in [45].

For evaluating the SENSEI prototype with respect to the second task of synopsis generation, we have set up an extrinsic evaluation task. The task aims at identifying if, and to what extent, the availability of automatically generated summaries may help QA supervisors in mining conversation types such as problematic calls. Focusing on problematic calls is important because it may potentially reduce the time-to-completion of tasks related with the supervision of call centre agents. At present a great number of calls need to be listened to and assessed in order to identify the potentially problematic ones as soon as they occur in the call centre. The design of this task is based on a focus group methodology, whose goals are the discovery of shared views among the participants, and the implications behind those views for the SENSEI speech prototype. The evaluation task requires that the group participants should be representative of the potential population of users of SENSEI speech prototype. In [48] we identified quality assurance and human resources professionals as end-users, and participants form that user group has been recruited for the focus group.

In the Table 5 we report the end-user comments (right column) that have emerged from the discussions for each question (left column). In that discussion we had four participants plus the moderator: participants A and B were QA

Table 5. Comments on focus group questions. A, B are QA supervisors and C is a quality assurance manager.

Question	Comment
How was your experience while using ACOF ?	A and B reported a positive experience
ACOF could highlight agent's behaviour?	A and B gave a positive answer
Did you agree with the ratings of the automatically filled ACOFs?	Most of the time
Do you expect that SENSEI ACOFs may help you in saving time in your job?	A, B, and C agreed
Do you think ACOFs could be enriched with evidence of the system decisions?	A and B gave positive answers
Usefulness of the synopses of the call?	A, B, and C think synopses might be useful
Why synopses could be useful?	All: To assess first call resolution and reasons for inbound calls
What is SENSEI potential added value for your job?	All: SENSEI system may allow to supervise a larger number of calls

supervisors, participant C was a quality assurance manager, and participant D was an HR specialist. As for the turn taking within the focus group, the conversations have been smooth and the participants have been collaborative.

In general, the focus group participants have found that the SENSEI results could be useful for their job because they would allow a larger number of calls to be monitored. They have also recommended that the automatic selection of problematic calls could be useful for partially overcoming the biases of human evaluation.

5.2 Social Media Use Case

In the case of reader comment summarization, identifying a user task poses challenges. This is because no one currently writes summaries of these comments as part of some larger task nor is there an obvious current user task setting in which summaries of reader comments would prove helpful. That said, our user study [48] has revealed considerable interest in such summaries and a wide set of user types and task settings where such summaries might play a useful role. One user task that could prove useful across end-user types, is that of automatically generating an overview of the key issues discussed in a reader comment set and the positions taken on these issues. We have constructed the following task-based evaluation motivated by this scenario. Further details may be found in [44,45]. To the best of our knowledge this is the first task-based evaluation protocol for reader comment summaries yet proposed.

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

1. Overview Questions: first, we ask participants to play the role of a user wanting to make sense of a comment conversation in a short period of time, e.g. a coffee break; we then provide users with a system and a topic (an article and comment set); allow a set time for reading over news and comment (e.g. 2 min) and then ask users to: (1) identify four main issues in the discussion and (2) characterise opinion on a given issue in a set time (e.g. 10 min) in accordance with our definitions.
2. Post task questionnaire: we ask participants to rate and compare the usefulness of the system(s) and system components in the context of completing the tasks, on a five point scale and include an option for written feedback.
3. Finally, in a guided group discussion we invite participants to comment on their experience during the tasks and on using the different systems/components.

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

A Pilot Evaluation. Participant responses to the overview questions are assessed manually. Assessors are given the source comments and the gold standard summaries (we select only articles which also appear in our gold standard for the extrinsic evaluation) and are asked to score written responses on a graded scale. The issues identified by participants in response to the overview questions are scored on a 4 point scale that takes account of criteria such as evidence/accuracy and clarity of expression. Characterisation of opinion is scored on a graded 6 point scale, based on criteria of coverage, representing quantities and accuracy. We analyze the free text and spoken responses gathered in the post task questionnaire and discussion using simple qualitative techniques. Data from the user ratings of the different systems/system components is summarised using simple statistics.

To carry out comparative evaluations of different systems we have developed a configurable interface with the following characteristics. It includes a baseline comment-only system, which presents threaded conversations in the way they typically appear in on-line news today, for example on *The Guardian* website. It takes as input comment clusters, labels for these clusters and summaries, which may be either extractive or abstractive and may contain links between sentences in the summary and the comment cluster that gave rise to the sentence. It offers two summary presentation modes: a text-based summary presentation mode and a graphical summary presentation mode. In the text-based mode the supplied summary and a textual representative of each cluster (e.g. a cluster label or

representative phrase or sentence) are displayed. The sentences in the summary, if links to clusters are provided, are clickable allowing the clusters underlying the sentences to be displayed. The textual representative of the clusters are also clickable allowing the comments in the cluster to be displayed.

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

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

S2. Included both the baseline functionality and sense-making components, consisting of a labelled pie chart indicating the relative size of comment clusters and a textual summary whose sentences were linked to underlying comment clusters. The clustering, cluster labelling and summarization outputs were produced by the top performing component combination described in Sect. 4.2 above, the ArticleLead-Sim-CentroidClosest-Comment-in-Thread system.

There were two different topics, each comprising a news article and an associated set of 100 comments. Each participant used each system and each topic exactly once. We provided a short training session including a system demo and guidelines on the overview scenario and tasks. We scored answers to the content questions using the metrics described above, aggregated ratings from the feedback questionnaire, and carried out a qualitative analysis of feedback from the group discussion. The three complementary sets of results allowed us to assess the protocol and to compare how, and to what extent, the different systems and system components helped users to complete the two content-related questions. While feedback on the usefulness of the sense-making technologies suggested more development was necessary if outputs were to help in such contexts, the general interface design and direction of the technology, as guided by the overview task, was approved of. The results also indicated that the protocol provides sufficient data to answer questions such as did different systems help with different content questions? Did one system help better overall? What features of the interface did users find most helpful in the task context? etc. A complete description of the methodology and evaluation task is given in [45].

6 Conclusion

The SENSEI project aims at taking a radically new approach at developing the technology for language summarization. We have selected a very relevant domain for the evaluation of the summarization technology: human conversations generated in contact centers and user comments on on-line news articles. By taking a *vertical* approach to the evaluation of the technology we have connected the end-users (e.g. customers or journalists) to the speech and natural language processing components and we expect to impact the efficacy of summary definition, generation and assessment. While improving the value of the

summary on the end-user task, we have shown that new research on semantic, para-semantic and discourse parsing has greatly contributed to the automatic generation of a novel type of summaries.

Acknowledgments. The research leading to these results has received funding from the European Union - Seventh Framework Program (FP7/2007-2013) under grant agreement no 610916.

The research reported in this paper would have not been possible without the key contributions of the researchers, engineers and professionals at the University of Trento, Aix-Marseille University, University of Sheffield, University of Essex, Websays and Teleperformance.

References

1. Boltuzic, F., Snajder, J.: Back up your stance: recognizing arguments in online discussions. In: Proceedings of the 1st Workshop on Argumentation Mining, Baltimore, MD, pp. 49–58 (2014)
2. Celli, F., Riccardi, G., Ghosh, A.: Corea: Italian news corpus with emotions and agreement. In: Proceedings of CLIC, Pisa (2014)
3. Erol, B., Lee, D.S., Hull, J.: Multimodal summarization of meeting recordings. In: 2003 International Conference on Multimedia and Expo, ICME 2003, Proceedings, vol. 3, pp. III–25. IEEE (2003)
4. Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., Mitsui, M.: Analyzing argumentative discourse units in online interactions. In: Proceedings of the 1st Workshop on Argumentation Mining, Baltimore, MD, pp. 39–48 (2014)
5. Giannakopoulos, G., Kubina, J., Conroy, J.M., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., Poesio, M.: Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In: Proceedings of SIGDIAL, Prague, pp. 270–274 (2015)
6. Gillick, D., Riedhammer, K., Favre, B., Hakkani-Tur, D.: A global optimization framework for meeting summarization. In: Proceedings of ICASSP 2009, Taiwan, pp. 4769–4772. IEEE (2009)
7. Kabadjov, M., Bechet, F., Favre, B., Kruschwitz, U., Poesio, M.: A coreference resolver for french spoken conversations (2015, submitted)
8. Kabadjov, M., Steinberger, J., Barker, E., Kruschwitz, U., Poesio, M.: Onforums: the shared task on online forum summarisation at multiling '15. In: Proceedings of FIRE (2015)
9. Kintsch, W., van Dijk, T.: Towards a model of discourse comprehension and production. Psychol. Rev. **85**, 363–394 (1978)
10. Lai, C., Renals, S.: Incorporating lexical and prosodic information at different levels for meeting summarization. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
11. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81 (2004)
12. Lindmark, D.: Methods for lean, precision-oriented, and targeted coreference resolution. Ph.D. thesis, Uppsala University (2012)
13. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: towards a functional theory of text organization. Text **8**(3), 243–281 (1988)

14. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge (2000)
15. Mehdad, Y., Carenini, G., Tompa, F.W., Ng, R.T.: Abstractive meeting summarization with entailment and fusion. In: Proceedings of the 14th European Workshop on Natural Language Generation, pp. 136–146 (2013)
16. Meyer, B.: The Organization of Prose and its Effects on Memory. North-Holland, Amsterdam (1975)
17. Murray, G.: Abstractive meeting summarization as a Markov decision process. In: Barbosa, D., Milios, E. (eds.) Canadian AI 2015. LNCS, vol. 9091, pp. 212–219. Springer, Heidelberg (2015)
18. Palau, R.M., Moens, M.F.: Argumentation mining. J. Artif. Intell. Law **19**(1), 1–22 (2011)
19. Poesio, M., Stevenson, R., Di Eugenio, B., Hitzeman, J.M.: Centering: a parametric theory and its instantiations. Comput. Linguist. **30**(3), 309–363 (2004)
20. Poesio, M., Stuckardt, R., Versley, Y.: Anaphora Resolution: Algorithms, Resources and Applications. Springer, Berlin (2016)
21. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The penn discourse treebank 2.0. In: Proceedings of LREC (2008)
22. Rodriguez, K.J., Delogu, F., Versley, Y., Stemle, E., Poesio, M.: Anaphoric annotation of wikipedia and blogs in the live memories corpus. In: Proceedings of LREC (poster) (2010)
23. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: Proceedings of COLING, pp. 1501–1510 (2014)
24. Steinberger, J., Kabadjov, M., Poesio, M., Pouliquen, B., Steinberger, R.: WB-JRC-UniTN's participation in TAC 2009: update summarization and AESOP tracks. In: Proceedings of TAC, Washington, November 2009
25. Steinberger, J., Poesio, M., Kabadjov, M., Jezek, K.: Two uses of anaphora resolution in summarization. Inf. Process. Manage. **43**(6), 1663–1680 (2007). Special Issue on Summarization
26. Stepanov, E.A., Riccardi, G., Bayer, A.O.: The unitn discourse parser in conll 2015 shared task: token-level sequence labeling with argument-specific models. In: Proceedings of CONLL - Shared Task, Bejing, pp. 25–31 (2015)
27. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
28. Taboada, M., Mann, W.C.: Applications of rhetorical structure theory. Discourse Stud. **8**(4), 567–588 (2006)
29. Uryupina, O., Moschitti, A., Poesio, M.: Bart goes multilingual: the unitn/essex submission to the conll-2012 shared task. In: Proceedings of the 15th CONLL: Shared Task. Association for Computational Linguistics, Korea, July 2012
30. Versley, Y., Ponzetto, S., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A.: Bart: a modular toolkit for coreference resolution. In: Proceedings of ACL, Demo Session, Columbus, OH, June 2008
31. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555 (2014)
32. Waibel, A., Bett, M., Finke, M., Stiefelhagen, R.: Meeting browser: tracking and summarizing meetings. In: Proceedings of the DARPA Broadcast News Workshop, pp. 281–286. Citeseer (1998)
33. Walker, M., Tree, J.F., Anand, P., Abbott, R., King, J.: A corpus for research on deliberation and debate. In: Proceedings of LREC (2012)

34. Wang, L., Cardie, C.: Focused meeting summarization via unsupervised relation extraction. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 304–313. Association for Computational Linguistics (2012)
35. Zechner, K.: Automatic summarization of open-domain multiparty dialogues in diverse genres. Comput. Linguist. **28**(4), 447–485 (2002)
36. Making Sense of Human-Human Conversations Data, Project FP7/2007-2013. http://www.sensei-conversation.eu
37. Celikyilmaz, A., Hakkani-Tur, D., Pasupat, P., Sarikaya, R.: Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. In: AAAI - Association for the Advancement of Artificial Intelligence (2015)
38. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2787–2795 (2013)
39. Chowdhury, A., Danieli, M., Riccardi, G.: The role of speakers and context in classifying competition in overlapping speech. In: Proceedings of INTERSPEECH, Dresden (2015)
40. Chowdhury, A., Danieli, M., Riccardi, G.: Annotating and categorizing competition in overlap speech. In: Proceedings of ICASSP, Brisbane (2015)
41. Riccardi, G., Stepanov, E., Chowdhury, A.: Discourse connective detection in spoken conversations. In: Proceedings of ICASSP, Shangai (2016)
42. Danieli, M., Riccardi, G., Alam, F.: Emotion unfolding and affective scenes: a case study in spoken conversations. In: Proceedings of ICMI Workshop on Representations and Modeling for Companion Systems, Seattle (2015)
43. Danieli, M., Balamurali, A.R., Stepanov, E., Favre, B., Bechet, F., Riccardi, G.: Summarizing behaviours: an experiment on the annotation of call-centre conversations. In: Proceedings of LREC, Portoroz (2016, to appear)
44. Barker, E., Funk, A., Paramita, M., Kurtic, E., Aker, A., Foster, J., Hepple, M., Gaizauskas, R.: What's the issue here?: task-based evaluation of reader comment summarization systems. In: Proceedings of LREC, Portoroz (2016, to appear)
45. Danieli, M., Barker, E. (eds.): Report on Intermediate Evaluation, SENSEI Deliverable D1.3, October 2015
46. Aker, A. (ed.): Report on Specification of Conversation Analysis and Summarization Outputs, SENSEI Deliverable D5.2, October 2015
47. Bayer, A.O., Riccardi, G.: Deep semantic encoding for language modeling. In: Proceedings of INTERSPEECH, Dresden (2015)
48. Danieli, M., Gaizauskas, R. (eds.): Report on Use Case Design and User Requirements, SENSEI Deliverable D1.2, October 2014
49. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
50. Khabiri, E., Caverlee, J., Hsu, C.-F.: Summarizing user-contributed comments. In: Proceedings of The Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011), pp. 534–537 (2011)
51. Ma, Z., Sun, A., Yuan, Q., Cong, G.: Topic-driven reader comments summarization. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), pp. 265–274 (2012)
52. Llewellyn, C., Grover, C., Oberlander, J.: Summarizing newspaper comments. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, pp. 599–602 (2014)

53. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space, CoRR, 1301.3781 (2013)
54. Murray, G., Carenini, G.: Summarizing spoken and written conversations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 773–782 (2008)
55. Sundermeyer, M., Ney, H., Schluter, R.: From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(3), 517–529 (2015)
56. Tafforeau, J., Artieres, T., Favre, B., Bechet, F.: Adapting lexical representation and OOV handling from written to spoken language with word embedding. In: Sixteenth Annual Conference of the International Speech Communication Association, INTERSPEECH (2015)
57. Uthus, D.C., Aha, D.W.: Plans toward automated chat summarization. In: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pp. 1–7 (2011)
58. Vukotic, V., Raymond, C., Gravier, G.: Is it time to switch to word embedding and recurrent neural networks for spoken language understanding? In: InterSpeech (2015)
59. Nenkova, A., McKeown, K.: Automatic summarization. Found. Trends Inf. Retrieval **5**(2–3), 103–233 (2011)
60. Carenini, G., Murray, G., Ng, R.: Methods for Mining and Summarizing Text Conversations. Morgan and Claypool Publishers, San Rafael (2011)
61. Byrd, R.J., Neff, M.S., Teiken, W., Park, Y., Cheng, K.S.F., Gates, S.C., Visweswariah, K.: Semi-automated logging of contact center telephone calls. In: Proceedings of CIKM (2008)
62. Higashinaka, R., Minami, Y., Nishikawa, H., Dohsaka, K., Meguro, T., Takahashi, S., Kikui, G.: Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues. In: Proceedings of COLING (2010)
63. Tamura, A., Ishikawa, K., Saikou, M., Tsuchida, M.: Extractive summarization method for contact center dialogues based on call logs. In: Proceedings of IJCNLP (2011)
64. Favre, B., Stepanov, E., Trione, J., Bechet, F., Riccardi, G.: Call centre conversation summarization: a pilot task at multiling. In: Sigdial (2015)
65. Oya, T., Mehdad, Y., Carenini, G., Ng, R.: A template-based abstractive meeting summarization: leveraging summary and source text relationships. In: Proceedings of International Conference on Natural Language Generation (INLG 2014) (2014)
66. Aker, A., Kurtic, E., Hepple, M., Gaizauskas, R., Di Fabbrizio, G.: Comment-to-article linking in the online news domain. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015), pp. 245–249 (2015)
67. Aker, A., Kurtic, E., Balamurali, A.R., Paramita, M., Barker, E., Hepple, M., Gaizauskas, R.: A graph-based approach to topic clustering for online comments to news. In: Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 15–29. Springer, Switzerland (2016)
68. Das, M.K., Bansal, T., Bhattacharyya, C.: Going beyond Corr-LDA for detecting specific comments on news & blogs. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 483–492 (2014)
69. Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out (2004)