Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

QA^{RT} : A System for Real-Time Holistic Quality Assurance for Contact Center Dialogues

Shourya Roy, Ragunathan Mariappan, Sandipan Dandapat, Saurabh Srivastava, Sainyam Galhotra and Balaji Peddamuthu
Xerox Research Centre India
Bangalore, India
{firstname.lastname@xerox.com}

Abstract

Quality assurance (QA) and customer satisfaction (C-Sat) analysis are two commonly used practices to measure goodness of dialogues between agents and customers in contact centers. The practices however have a few shortcomings. QA puts sole emphasis on agents' organizational compliance aspect whereas C-Sat attempts to measure customers' satisfaction only based on post-dialogue surveys. As a result, outcome of independent QA and C-Sat analysis may not always be in correspondence. Secondly, both processes are retrospective in nature and hence, evidences of bad past dialogues (and consequently bad customer experiences) can only be found after hours or days or weeks depending on their periodicity. Finally, human intensive nature of these practices lead to time and cost overhead while being able to analyze only a small fraction of dialogues. In this paper, we introduce an automatic real-time quality assurance system for contact centers - QA^{RT} (pronounced *cart*). QA^{RT} performs multi-faceted analysis on dialogue utterances, as they happen, using sophisticated statistical and rule-based natural language processing (NLP) techniques. It covers various aspects inspired by today's QA and C-Sat practices as well as introduces novel incremental dialogue summarization capability. QA^{RT} front-end is an interactive dashboard providing views of ongoing dialogues at different granularity enabling agents' supervisors to monitor and take corrective actions as needed. We demonstrate effectiveness of different back-end modules as well as the overall system by experimental results on a real-life contact center chat dataset.

Introduction

Contact center is a general term for help desks, information lines and customer service centers. They provide dialogue (both voice and online chat) and email-based support to solve product and services-related issues, queries, and requests. Two key drivers of contact center industry are cost reduction and service quality improvement. Exploiting cost arbitrage through outsourcing and bringing in automation for (parts of) service delivery processes such as agent assistance tools (Padmanabhan and Kummamuru 2007; Byrd et al. 2008; Marom and Zukerman 2009) have been companies' strategy towards the first.

Providing highest quality of service leading to satisfied customers is the other key objective of contact centers. Two of the most commonly employed practices towards that are *Quality Assurance* (QA) and *Customer Satisfaction analysis* (C-Sat). QA process, primarily involving *dialogue scrutiny* (offline) and to some extent *dialogue monitoring* (live), is about measuring quality against a set of targets defined by the organization.[1] Supervisors of contact center agents are expected to rate agents' performances on a variety of metrics to measure how compliant and effective they have been (intrigued readers may make a forward reference to Table 5 to see some examples). C-Sat, on the other hand, is about analyzing customers' post interaction feedback to identify drivers for (dis)satisfaction. Manual customer satisfaction surveys are conducted via telephonic interviews, mail-in or electronic forms, where customers are asked to evaluate various aspects of their interaction on a 5-point Likert scale (Likert 1932) and subsequently responses are analyzed. While a number of research articles have been written about correlation between these two practices and their comparative usefulness (Kantsperger and Kunz 2005; Rafaeli, Ziklik, and Doucet 2008), both remain widely popular in the industry. Contact centers employ specialized people to conduct these processes periodically and typically independently to identify actionable areas of improvement in service delivery (e.g. leading to customized agent training).

Both manual QA and C-Sat as practiced today have several shortcomings. Firstly, they reveal outcomes of dialogues and associated reasons always in hindsight. While live dialogue monitoring is encouraged but owing to their human time and effort intensive nature, it is less common than their offline counterpart. Secondly, owing to their different focus of analysis - customers' post-interaction feedback in C-Sat versus best practices mandated by the organization in QA, it is not uncommon to have conflicting outcome from these two processes. An agent and her supervisor may feel that in a dialogue everything was done perfectly but still customer's feedback could be negative. Thirdly, only a small fraction of contact center workforce are responsible for QA and C-Sat (Godbole and Roy 2008c) processes, hence they can analyze only a small sample of total interactions thereby

[1] We use the term *dialogue* to refer to two-party contact center conversations between customers and agents.

missing out on the most.

In this paper we introduce QA^{RT}, a system for quality assurance in real-time in contact centers. QA^{RT} performs holistic multi-faceted analysis on each and every utterance[2] made by customers and agents by bringing together and automating various aspects of manual QA and C-Sat processes. For instance, *Organizational Compliance* and *Conversational Characteristics* facets spot occurrences of deviations from prescribed or expected QA metrics. *Customer Behavior* facet acts as a real-time proxy for C-Sat feedback from customers by identifying sentiments (e.g. positive and negative) and emotions (e.g. angry, sad, satisfied) on utterances. An example dialogue from a telecommunication contact center and associated issues can be seen in Table 1. We have developed novel features and techniques in natural language processing (NLP) and built on the state-of-the-art machine learning to extract relevant information from utterances. Beyond modules for real-time automation of existing processes, QA^{RT} introduces two novel components. An incremental dialogue summarizer which generates actionable summaries for supervisors to quickly gather context and possibly intercept a dialogue faster. Secondly, it provides an interactive dashboard showing real-time status of ongoing dialogues at various granularity. It enables supervisors to obtain glanceable views and on demand details about potentially problematic dialogues in real-time.

QA^{RT} offers several advantages. First and foremost, to the best of our knowledge QA^{RT} is the first end-to-end, from utterances to visualization, real-time QA system. It performs holistic analysis by extending and enriching existing QA practices and by including behavioral aspects of customer satisfaction. Secondly, QA^{RT} is real-time and bad customer experiences can be prevented based on instantaneous status which is a significant advantage over current retrospective processes. Actionable real-time summaries of dialogues makes interventions seamless. Thirdly, it is automatic and thereby can handle large number of interactions simultaneously and thereby enabling contact centers to go beyond sampling based QA. Results from experiments on a real-life corpora of 188 dialogues from a telecommunication contact center bring out efficacy of various back-end NLP modules as well as the overall system.

Organization of the paper: After briefly reviewing relevant prior research in the next section, we provide technical details of back-end NLP modules to extract different features from utterances relevant for QA. Subsequently, we cover an incremental dialogue summarization technique. Next we describe the QA^{RT} front-end dashboard and how it brings together outcome of different back-end modules. Finally we describe the experimentation methodology and results to compare QA^{RT} against sampling based QA process in a controlled environment.

[2]An utterance is a contiguous sequence of words from one of the parties in a dialogue. We use the words *turn* and *utterance* interchangeably in this paper.

Related Work

Service quality and customer satisfaction in contact centers have seen large body of qualitative and more recently, computational work. Qualitative studies have indicated correlation of varied strengths between agents' customer orientations with customer evaluations of service quality (Rafaeli, Ziklik, and Doucet 2008) and customers' affective commitment and loyalty (Dean 2007). On the computational side, several techniques have been developed, primarily from industrial researchers, based on call logs, transcribed conversations, chats providing various types of analytics and insights such as agent assistance (Padmanabhan and Kummamuru 2007; Byrd et al. 2008; Marom and Zukerman 2009), knowledge generation (Roy and Subramaniam 2006; Lee et al. 2009), discovering business insights (Takeuchi et al. 2009; Cailliau and Cavet 2013). Among these automatic C-Sat is the most relevant for this work which we review next.

Godbole and Roy (2008a; 2008b; 2008c) developed a text classification based system for categorizing C-Sat comments into 31 classes. They extended scope for C-Sat analysis to automatically discover *reason codes* from customer comments. They also provided a visual interface for offline reporting and querying. However, their system by design could automate only the existing retrospective C-Sat analysis process and there was no real-time aspect of their work. Park and Gates (2007) identified several interesting features from 115 transcribed conversations for real-time C-Sat analysis on a 5-point Likert scale. While they were the first to introduce and experimentally justify the notion of (near) real-time C-Sat analysis, their work is limited by the scope of C-Sat analysis process. Their approach could provide, at best, a point estimate of C-Sat score with no actionable insights about what is (not) working well in a contact center. Li and Chen (2010) mined product and service reviews to automatically generate customer satisfaction surveys. Mishne et al. (2005) described a system for call center analysis on monitoring based on manually transcribed conversations. Their reported results are only on problem identification and detecting off-topic segments. Zweig et. al. (2006) employed an offline system using rule-based and maximum entropy classifier for identifying bad calls based on a predefined list of questions. Overall, we recognized a void towards an end-to-end real-time quality assurance system capable of providing actionable insights and interactivity. QA^{RT} is designed and developed to fill that void.

Facets and Features

In this section, we describe various features and facets (collections of similar features) relevant for conducting real-time quality monitoring. Customer and agent utterances are passed through these modules to extract various features. For each feature, we present relevant experimental results to demonstrate efficacy of our techniques and associated attributes.

12:24:29	AGENT:	Hi there, thanks for contacting Mobile Product Support. My name is AGENT, how can I assist you today?
12:25:05	CUST:	i just swapped my MODEL-1 phone for a MODEL-2. how do i get to my voicemail.
12:26:48	CUST:	is anyone there **(Customer Behavior: Customer is unhappy)**
12:29:09	AGENT:	You can access your voicemail by pressing and holding on the 1 key **(Conversational characteristic: Agent's delayed/unusual response time)**
12:29:44	CUST:	i did that and it said that it wasnt set up
12:29:59	AGENT:	Are you using the same carrier? **(Organizational compliance: Agent didn't apologize)**
12:30:19	CUST:	no i switched from PROVIDER-1 to PROVIDER-2
...	...	...

Table 1: An example dialogue from our dataset

Customer Behavior

Experience of a customer can be best understood and potentially measured by her expressed views and opinions as well as her feeling expressed during a dialogue. In NLP, identifying the former is the task of *sentiment categorization* (Turney 2002) whereas the latter is that of *emotion categorization* (Krcadinac et al. 2013). In QA^{RT}, we focus on identifying emotions (such as *sad*, *apologetic* or *angry*) of customers from their utterances and obtain sentiment categories (such as *positive*, *negative*) as a by-product. We believe, identified emotions capture customers' mental state during dialogues and act as proxies of subsequent C-Sat feedback. While this may not be always true but qualitatively we found evidences for such correspondence by seeing through a sample of past dialogues and associated C-Sat feedback. We describe key aspects of the module below:

Tag-set: We started by creating a emotion *tag-set* for contact center dialogues. While there have been past work towards defining emotion tag-sets (Ekman and Keltner 1970; Plutchik 2003), we could not adopt those in entirety as some tags did not make sense in our context (e.g. *fear*) and we needed finer categories at other places (e.g. *agreement* and *disagreement*). Towards that, we created a contact center specific 8-class emotion tag-set viz. *Happiness* (Ha), *Assurance* (As), *Agreement* (Ag), *Courteousness* (Co), *Apology* (Ap), *Unhappiness* (Uh), *Disagreement* (Di) and *No-Emotion* (Ne). Additionally, categories {Ha, As, Ag, Co} are considered as positive, {Ap, Uh, Di} as negative and {Ne} as neutral sentiment to obtain sentiment categorization for utterances.

Attributes: We use conversational meta-attributes along with content-based ones for representing utterances to perform emotion categorization. Two sets of attributes are shown in Table 2. While both sets of attributes have been designed to capture conversational aspect of the problem, particularly meta-attributes exploit sequence and inter-relationship between utterances of agents and customers.

Technique: The emotion categorization technique is composed of two main tasks – (i) splitting a turn into parts which are homogeneous with respect to emotion and (ii) assigning emotion categories to turns. We observed that many customer turns are multi-part with respect to emotions. Consider the example below:

> CUST: ⟨Hi AGENT!⟩ ⟨I am having a problem with my PHONE. Whenever I connect to my computer via usb, it will only charge and not recognize the SD card. If I take out the card and put it in an adapter it works.⟩ ⟨ Can you help me?⟩

This customer turn has three segments which are demarcated with angular brackets. The first one is a courteous greeting (Co), middle one describes a problem (Uh) and finally a request which is neutral (Ne). Assigning a single emotion label to the turn would have been misleading. We note that most multi-emotion turns are also multi-sentence long and typically, within a sentence typically emotion does not change. We implemented a boundary detection module based on simple heuristics which identified sentence boundaries with 86% accuracy on our corpus of dialogues.

Subsequently, given a sequence of sentences $S = s_1, s_2, \ldots, s_n$ represented using attributes described above, we want to find the corresponding sequence of emotions $e = e_1, e_2, \ldots e_n$ from the set of emotion tag-set. We modeled this as a sequential labeling problem and applied Conditional Random Field (CRF) (Lafferty, McCallum, and Pereira 2001). The conditional dependency between e_i and S is defined through the attribute functions of the form $f(i, e_{i-1}, e_i, s)$. We have used open source implementation CRF++ [3] to train and test our models. Result on the corpus of 188 dialogues is shown in Table 3 using leave-one-out cross-validation. The best accuracy of customers' emotion detection obtained is 62.4% which is reasonable considering this was a 8-class classification task and evidences of emotions in dialogues were relatively infrequent. Additionally, inter annotator human agreement (between two annotators) measured by Cohen's Kappa (Cohen 1960) score was 78% which shows the task is non-trivial even for humans. We also report the model's performance on agent utterances where it gives higher accuracy due to their typically homogeneous nature.

Conversational Characteristics

Call center conversations exhibit certain regularities with respect to their structure and flow; we call them *conversational characteristics*. Dialogues deviating from these characteristics are expected to be flagged as anomalous during QA process. Detecting such deviations in manual QA process is difficult as one needs to consider multiple factors and their expected behavior. In this section, we describe how we automatically detect such deviations with respect to conversation structure and certain content independent features directly from utterances in a real-time fashion.

[3] http://taku910.github.io/crfpp/

Content-based Attributes	
BOW	bag of content words whose term frequency ≥ 5
Polar word	number of positive and negative words in a turn (using Synesketch (Krcadinac et al. 2013) lexicon)
Emoticon	number of positive and negative emoticons
Subsentential	number of items separated by punctuation
Lex	number of subsentential with lexical items: *but*, *too* or any *negations*
Prev tag	emotion/Sentiment category of the previous turn.
Meta Attributes	
A/C label	boolean feature indicating Agent(A) or Customer(C) turn
Delay	time difference between two consecutive turns
Position	turn sequence number in a dialogue
Length	number of words in a turn
Turn Type	turn segments indicating *general statement*, *query*, *problem statement*, *response to a query*

Table 2: Different attributes used for emotion categorization

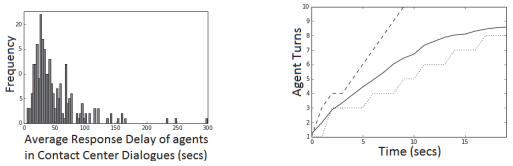
	Customers'		Agents'	
Features	Emotion	Sentiment	Emotion	Sentiment
Baseline (BOW)	40.5	61.6	75.5	80.9
Content	44.7	66.1	79.3	82.1
Content+Meta	**62.4**	**69.5**	**79.8**	**82.0**

Table 3: Emotion and sentiment detection accuracy (in %)

	Segment level				Overall
	GI	PD	PR	CL	
SVM+Rules	100	47	98	30	80.3
CRF+Rules	100	70	91	56	83.9
K-means+Rules	100	50	91	38	82.8

Table 4: Accuracy of conversation structure detection

Conversation Structure: Contact center dialogues typically follow well defined structure in terms of an ordered sequence of states. In telecommunication domain, we observed that commonly observed sequence of states is {*greetings and introduction* (GI) $\rightarrow$ *problem description* (PD) $\rightarrow$ *problem resolution* (PS) $\rightarrow$ *closure* (CL)}. Detecting deviations from this prescribed order and identifying missing states are important aspects of manual QA process. In QA^{RT}, we apply both supervised and unsupervised techniques to detect any such discrepancy in real-time by grouping turns into one of the four categories (GI, PD, PS and CL). Firstly, we note that this can be modeled as a sequence labeling problem and hence, we applied CRF with a subset of attributes used for emotion categorization viz. BOW and turn number. As baseline, we used a standard classification model (Support Vector Machine (SVM) (Cortes and Vapnik 1995)) which assumes turns to be independent and does not use sequential information. In contrast to supervised CRF and SVM models, we also applied unsupervised K-means ($k = 4$) clustering algorithm on utterances to group them into 4 groups. The clustering algorithm uses inter-turn similarities (based on cosine similarity on BOW representation of turns) and some domain specific attributes such as inverse of the difference between two turn numbers. Furthermore, we created a few hand-written rules to improve accuracy of certain categories which were used across all the techniques. For example, a turn with "Thanks for contacting" and "have a nice day" is bound to be in $closure$ category. Table 4 shows comparative results of these techniques on our dialogue corpus. Although accuracy obtained by CRF is marginally better than the K-means algorithm, the latter does not require labeled data thereby needed no human supervision.



(a) (b)

Figure 1: (a) Distribution of average response delays (b) Deviating and compliant dialogues with respect to the feature *Number of Agent Turns*. The middle curve shows average behavior, the dotted-line curve shows a compliant dialogue whereas the dashed-line curve shows a deviating one (after 5^{th} turn).

Conversation Features: Apart from structural regularities, call center conversations also exhibit regularities with respect to some content independent features e.g. *number of turns taken by an agent (or customer) per minute*, *average inter-turn delays* and so on. For example, distribution of *agents' average response delay* in our corpus as shown in Figure 1(a) is roughly centered around 25 seconds on the x-axis.

These features are represented as sequences of discrete real numbers and their temporal averages are computed over duration of dialogues. During an ongoing dialogue, the task is to detect whether the same feature is deviating significantly from the average (illustration in Figure 1b). We employ the classical paired 2-sample *t*-test (Diggle et al. 2002) to detect such anomalous behaviors as following:

- Compute mean feature value up to the j^{th} minute as $\overline{a}_j$
- For every new dialogue
  - compute the same feature value up to the j^{th} minute as

x_j and their difference as $d_j = x_j - \overline{a}_j$

- compute 2-sample t-statistic till the j^{th} minute as $\frac{(\overline{d})}{SE(\overline{d})}$, where $(\overline{d})$ is the *mean* difference and $SE(\overline{d})$ is the *standard error* of mean difference.
- Under the null hypothesis, this statistic follows a t-distribution with $(j-1)$ degrees of freedom. If the p-value for the paired t-statistic from the t-distribution table is more, then the null hypothesis is rejected i.e. up to the j^{th} minute the new dialogue is deviating significantly from past average.
- convert t-statistic output for each period as a sequence of boolean values e.g. $F, F, T, T, F, \ldots$ indicating whether up to the j^{th} utterance were found to be deviating from past dialogues. We call the entire dialogue deviating with respect to the particular feature on observing a sequence of k anomalous periods. Empirically we found $k = 3$ gave us the best result.

We tested performance of the aforementioned approach to detect deviating dialogues with respect to agents' response delay on a part of our corpus. The training set were dialogues about *mobile display problems* which were considered as past conversations to compute average feature values. The test set were containing dialogues about both *network issues* (which should be identified as deviating or true positives) and *display problems* (true negatives). The precision and recall were found to be 67% and 63% respectively. The precision (fraction of predicted *network issues* problems which were correct) and recall (fraction of actual *network issues* problems which were predicted correctly) were found to be 67% and 63% respectively.

Organizational Compliance

Contact centers employ certain best practices towards ensuring uniform, courteous and effective interaction style across all agents. Some example best pratices are shown in Table 5. Today's QA process involves retrospective dialogue scrutiny to identify instances where one or more of these were not followed. We folded in this process as an automatic module in QA^{RT} and thereby enabling compliance checking for agent utterances in real-time. From a list of about 30 organizational guidelines provided to us by domain experts, we noted that most can be validated by applying regular expression like rules on agent utterances. We designed a rule-base consisting of these rules and other information such as scope, frequency and order of applications. For example, the rule for the first guideline shown in Table 5 is $t.isFirstTurn(Agent) \wedge t.startsWith(\text{“}Welcome\ to\ WeCare\ helpdesk.\text{”})$. We implemented this module as a customized rule-engine for easy portability but more sophisticated rule engine such as RIPPER (Cohen 1995) can be used as well. Some important aspects of the rule-base are listed below:

- Rules are organized in the same order in which they are expected be applied. For example the rule corresponding to `Did the agent assure the customer?` will be applied before the rule for `Did the agent perform required security check?`.
- Most rules are applicable only within specific states(s) as identified in conversation structure. We define this as *scope* of a rule. Few rules such as the one for `Did the agent interrupted or talked over the customer?` has however scope over entire dialogues.
- A rule can be optional or mandatory. A mandatory rule may appear one or more number of times in a dialogue.

Dialogue Summarization

The facets and features described above are useful to identify issues (e.g. angry emotion, unusual delay by agent) in contact center dialogues efficiently. However in practice to understand context of these issues supervisors would need to go through long transcripts. For example, a customer is exhibiting angry emotion because *provided resolution did not work* or an agent is showing anomalous behavior because *she is taking time to find the resolution*. To enable supervisors to obtain context at-a-glance, we introduced a novel task-oriented incremental dialogue summarization module in QA^{RT}(Figure 2c). Dialogue summaries consist of two parts – (i) an information template and (ii) dialogue excerpts in support of extracted information.

Template-Based Abstract: Certain information in a telecommunication contact center such as device name, problem type, etc. are essential to understand context of dialogues (Table 6). We have designed a template to present these information which gets incrementally filled up as conversations proceed. We use a combination of dictionary look-ups (for device names) and Random Forest classifiers with Bag-of-Word features (Problem Type and Problem Resolution Type) for information extraction from dialogues.

Summary Excerpts: The other part of the summarizer excerpts key phrases from dialogues in support of template entries. Manually analyzing tens of dialogues we observed that Parts Of Speech (POS) tags of such key phrases have some commonalities. $< VBZ >< VB.* >$ and $< VP >< IN > *$ are two such POS patterns.[4] We extracted actual phrases corresponding to these patterns to represent them using attributes described in Table 7 and annotated as 1 or 0 indicating whether or not they were key phrases. A classifier is then trained to predict whether a phrase to be extracted as a summary excerpt or not. We tried multiple classifiers and found Random Forest gave the best accuracy 85.9%.

Interactive Visualization

QA^{RT} provides a rich and interactive dashboard to visualize outcome of feature analysis and summarization module in a real-time fashion. The dashboard is another novel contribution as most prior work in this domain did not go beyond data analysis techniques to make them consumable by contact center operations people. It offers different views of ongoing dialogues and extracted information as well as interactional ability to slice and dice, roll-up and down. The

[4]Expansion of POS tags can be found at Penn Treebank project http://bit.ly/1gwbird

Category	Sample guidelines and Examples
Intro	– Was standard greeting made? *(Welcome to WeCare.)* – Was customer name used? *(Hello John. How are you doing today?)*
Soft skill	– Did the agent interrupt the customer? – Did the agent assure the customer? *(I can help you to make it working.)*
Technical	– Did the agent perform required security check?
Closure	– Did the agent provide reference number? *(Ref. no. for this chat is AH680.)* – Did the agent close appropriately *(Thank you for contacting WeCare.)*

Table 5: Sample agent compliance guidelines for contact center dialogues

Attribute	Description	Examples
Status	Capturing the current state of a dialogue in few words	problem description, resolution
Device	Device of interest in a dialogue	htc, iphone.
Problem	The problem reported	software issue, display issue
Problem Resolution	The problem resolution being advocated	hard reset, soft reset

Table 6: Summarization template attributes, their descriptions and examples in a telecommunication contact center

Features	Description
BoW	
Contextual	
LOP	Length of phrase extracted by PoS pattern
LOT	Length of the turn containing the phrase
GNLOT	LOT divided by the max LOT till now
GNLOP	LOP divided by global max LOP till now
LNLOP	LOP divided by max LOP observed in that same turn
Conversational	
TI	Turn Index(number) containing the phrase
TB	Time from the beginning of chat to the Turn containing the phrase
TP	Time from the previous turn to the turn containing the phrase.
IA	Is it an Agent's turn 0 or 1 yes or no
PP	Position of phrase in the turn (start index)
IQ	Is the turn a question 0 or 1 yes or no

Table 7: Features used to classify summary excerpts

dashboard is designed to provide quick at-a-glance views and obtain various levels of details on demand, instead of overloading by providing details of all dialogues at the same level.

Visualization and graphical representation in the dashboard follow Gestalts laws (Hartmann 1935) towards minimizing cognitive load. The primary view (Figure 2a) provides a top-level status of ongoing dialogues based on different facets. In the grid-structure, *good* dialogues are represented in spectrum of blue color whereas problematic ones (angry emotion, lack of compliance) are represented in spectrum of red color. This visualization is focused to enable a supervisor to easily monitor the overall summary of the ongoing chats, in parallel with monitoring of filtered, group, or even individual chats and features. If everything is going fine then no action is needed, alternatively they can drill down to a *red* dialogue by clicking on the corresponding grid cell which takes them to the features view pane. For example, Figure 2b shows detail view of *Organizational Compliance* and *Conversation Structure* features. If they want further details of the dialogue, presumably to intercept it, the summarization view (Figure 2c) brings up an actionable summary. Left panel shows the standard template with extracted information from the dialogue while the right panel shows corresponding summary excerpts highlighted.

Experiments

In this section, we present experimental results comparing QA^{RT} with manual QA process. We selected a real-world dataset of 188 contact center dialogues about mobile phone service providers and manually annotated instances of quality issues such as dialogues showing negative customer emotion, non-compliant or deviating with respect to conversation features (Table 8).

Facets	Issue description	#issues
Customer behavior (CB)	Negative sentiment	3
	Unhappy emotion	9
Conversation features (CF)	Anomalous characteristics	5
Organizational compliance (OC)	Missed to greet	1
	Missed to apologize	3
	Missed to assure	4
	Non-compliant closing	1
	Prohibited words usage	2

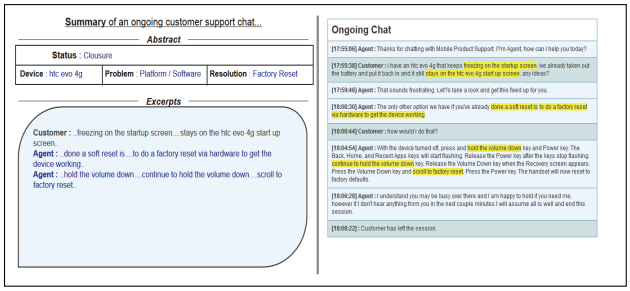
Table 8: Details of potential quality issues in 188 contact center dialogues in telecommunication domain

Each utterance comes with a timestamp (as shown in the dialogue in Table 1) which we used to simulate real-time streaming of dialogues. A confederate playing the role of a contact center quality analyst was asked to identify the issues across various facets with (i) the aid of QA^{RT}, against the baseline of (ii) conventional sampling based manual examination of chats. For (ii) we selected sampling rate to be 10% which is higher than typical single digit sampling of real QA process. We measured the following metrics:

- **Precision and recall**: Precision and recall are respectively defined as fraction of correctly reported issues and frac-

a: Snapshot view

b: Features view



c: Dialogue summary view

Figure 2: Different views of QA^{RT} dashboard (best viewed in color and at high zoom level)

Facets			P	R	Ry	T (s)
OC		QA^{RT}	0.81	1.0	1	2.2
		QA-M	0.69	1.0	(0.52, 0.08, 0.01)	1150
CB	Emo	QA^{RT}	0.61	0.72	1	5.6
		QA-M	0.90	0.67	$(0.63, 10^{-3}, 10^{-5})$	1280
	Senti	QA^{RT}	0.78	0.37	1	-
		QA-M	0.77	0.53	(0.3, 0.3, 0.1)	-
CF		QA^{RT}	0.73	0.33	1	3.4
		QA-M	0.0	0.0	(0.4, 0.4, 0.1)	502
RS		QA^{RT}	0.69	0.52	-	6.8
		QA-M	0.52	0.29	-	564

Table 9: Experimental results (P: precision, R: recall, Ry: reliability, T: time taken in seconds, RS:real-time summarization) comparing QA^{RT} and manual QA (QA-M). Reliability of QA-M is shown by tuples (-all, -at least 1, -same). Time taken for *Senti* is not shown as it is a by-product of *Emo*.

tion of issues in the dataset which are reported by the confederate. For sampling based manual examination of chats, we report precision and recall numbers based on number of issues present in the chosen sample.

- **Reliability**: This is the fraction of the issues which were present in the dataset looked at by the confederate. It measures how likely the sample will contain all/at least-one/the-same issues present in the dataset (in other words, whether the sample is reliable). For QA^{RT} this is always 1 as it analyzes all dialogues.

- **Time taken**: Time taken to perform a QA task.

Results: Table 9 shows the experimental results. Overall the confederate could achieve similar or better performance when enabled with QA^{RT} in almost all the facets while taking significantly less time. For features such as *conversation features*, QA^{RT} shows significant better performance as computing deviations is computationally intensive task. Low reliability for manual process demonstrates unreliability in achieving the precision and recall reported. In general, for such finding-few-needles-in-a-haystack task, sampling based approach could be misleading.

Conclusion

In this paper, we introduced QA^{RT}, an end-to-end real-time quality assurance system for contact centers. Beyond automating standard manual QA process, it introduces novel capabilities such as incremental dialogue summarization. We demonstrated its benefits experimentally by comparing against manual QA. As future work, we intend to extend QA^{RT} for spoken conversations based on transcribed speech as well as acoustic features.

References

Byrd, R. J.; Neff, M. S.; Teiken, W.; Park, Y.; Cheng, K.-S. F.; Gates, S. C.; and Visweswariah, K. 2008. Semi-automated logging of contact center telephone calls. In *Proc. of the 17th ACM conference on Information and knowledge management*, 133–142. ACM.

Cailliau, F., and Cavet, A. 2013. Mining automatic speech transcripts for the retrieval of problematic calls. In *Computational Linguistics and Intelligent Text Processing*. Springer. 83–95.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Cohen, W. W. 1995. Fast effective rule induction. In *Proc. of the 12th intl. conference on machine learning*, 115–123.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Mach. Learn.* 20(3):273–297.

Dean, A. M. 2007. The impact of the customer orientation of call center employees on customers' affective commitment and loyalty. *Journal of Service Research* 10(2):161–173.

Diggle, P.; Heagerty, P.; Liang, K.-Y.; and Zeger, S. 2002. *Analysis of longitudinal data*. Oxford University Press.

Ekman, P., and Keltner, D. 1970. Universal facial expressions of emotion. *California Mental Health Research Digest* 8(4):151–158.

Godbole, S., and Roy, S. 2008a. An integrated system for automatic customer satisfaction analysis in the services industry. In *Proc. of the 14th SIGKDD intl. conference on Knowledge discovery and data mining*, 1073–1076. ACM.

Godbole, S., and Roy, S. 2008b. Text classification, business intelligence, and interactivity: automating c-sat analysis for services industry. In *Proc. of the 14th SIGKDD intl. conference on Knowledge discovery and data mining*, 911–919. ACM.

Godbole, S., and Roy, S. 2008c. Text to intelligence: Building and deploying a text mining solution in the services industry for customer satisfaction analysis. In *Services Computing, 2008.*, volume 2, 441–448. IEEE.

Hartmann, G. W. 1935. Gestalt psychology: A survey of facts and principles.

Kantsperger, R., and Kunz, W. H. 2005. Managing overall service quality in customer care centers: Empirical findings of a multi-perspective approach. *Intl Journal of Service Industry Management* 16(2):135–151.

Krcadinac, U.; Pasquier, P.; Jovanovic, J.; and Devedzic, V. 2013. Synesketch: An open source library for sentence-based emotion recognition. *Affective Computing, IEEE Transactions on* 4(3):312–325.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the Eighteenth Intl. Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Lee, C.; Jung, S.; Kim, K.; and Lee, G. G. 2009. Automatic agenda graph construction from human-human dialogs using clustering method. In *Proc. of HLT: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-Short '09, 89–92. Stroudsburg, PA: ACL.

Li, S., and Chen, Z. 2010. Exploiting web reviews for generating customer service surveys. In *Proc. of the 2nd Intl. Workshop on Search and Mining User-generated Contents*, SMUC '10, 53–62.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22(140):1–55.

Marom, Y., and Zukerman, I. 2009. An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Comput. Linguist.* 35(4):597–635.

Mishne, G.; Carmel, D.; Hoory, R.; Roytman, A.; and Soffer, A. 2005. Automatic analysis of call-center conversations. In *Proc. of the 14th ACM Inl. Conference on Information and Knowledge Management*, CIKM '05, 453–459.

Padmanabhan, D., and Kummamuru, K. 2007. Mining conversational text for procedures with applications in contact centers. *International Journal of Document Analysis and Recognition (IJDAR)* 10(3-4):227–238.

Park, Y. 2007. Automatic call section segmentation for contact-center calls. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, 117–126. New York, NY, USA: ACM.

Plutchik, R. 2003. *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association.

Rafaeli, A.; Ziklik, L.; and Doucet, L. 2008. The impact of call center employees' customer orientation behaviors on service quality. *Journal of Service Research* 10(3):239–255.

Roy, S., and Subramaniam, L. V. 2006. Automatic generation of domain models for call centers from noisy transcriptions. In *Proc. of the 21st Intl. Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, 737–744.

Takeuchi, H.; Subramaniam, L. V.; Nasukawa, T.; and Roy, S. 2009. Getting insights from the voices of customers: Conversation mining at a contact center. *Inf. Sci.* 179(11):1584–1591.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 417–424. Association for Computational Linguistics.

Zweig, G.; Siohan, O.; Saon, G.; Ramabhadran, B.; Povey, D.; Mangu, L.; and Kingsbury, B. 2006. Automated quality monitoring for call centers using speech and nlp technologies. In *Proc. of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL-Demonstrations '06, 292–295.