Interaction Mining: the new frontier of Customer Interaction Analytics

Vincenzo Pallotta[1], Rodolfo Delmonte[2]

University of Business and International Studies
Av. Blanc, 46 1202 Geneva, Switzerland
vincenzo.pallotta@gmail.com

Department of Computational Linguistics
University 'Ca Foscari Venice, Italy.
delmont@unive.it

Abstract In this paper, we present our solution for argumentative analysis of call center conversations in order to provide useful insights for enhancing Customer Interaction Analytics to a level that will enable more qualitative metrics and *key performance indicators* (KPIs) beyond the standard approach used in Customer Interaction Analytics. These metrics rely on understanding the dynamics of conversations by highlighting the way participants discuss about topics. By doing that we can detect relevant situations such as social behaviors, controversial topics, customer oriented behaviors, and also predict customer satisfaction.

1 Introduction

Call centers data represent a valuable asset for companies, but it is often underexploited for business purposes. By call center data we mean all information that can be gathered from recording calls between representatives (or agents) and customers during their interactions in call centers. These interactions can happen over multiple different channels including telephone, instant messaging, email, web forms, etc. Some information can be collected without looking at the content of the interaction, by simply logging the system used for carrying the conversation. For example, in call centers, calls duration or number of handled calls can be measured by software for telephony communication. We call these measures standard call center Key Performance Indicators (KPIs). With standard KPIs, only limited analytics can be done. Of course, one can aggregate these measures over several dimensions of other meta-data such as agents, customers, regions, time, and queues. However, this only provides a partial understanding of the call center performance and no information whatsoever is collected about what is going on within the interaction.

Customer Interaction Analytics is aimed at solving the above issue by enabling tapping into the content of conversations. The technology for Customer Interaction Analytics is still in its infancy and related commercial products have not yet achieved maturity. This is due to two main factors: i) it is highly dependent on quality of speech recognition technology and ii) it is mostly based on text-based content analysis.

We believe that text-based content analysis approaches are highly sensitive to input quality and that conversational input is fundamentally different than text. Therefore, conversations should be treated differently. Natural Language Processing (NLP) technology needs be adapted to and robust enough to deal with the conversational domain in order to achieve acceptable performance. Moreover, the level of analysis of conversation cannot be set to semantics only. It must consider the purpose of language in its context, i.e., pragmatics.

Our approach to Customer Interaction Analytics is based on Interaction Mining. Interaction Mining is a new research field aimed at extracting useful information from conversations. In contrast to Text Mining (Feldman and Sanger 2006), Interaction Mining is more robust, tailored for the conversational domain, and slanted towards *pragmatic* and *discourse* analysis. We applied our technology for pragmatic analysis of natural language to a corpus of call center conversations and shown how this analysis can deal with the situations we mentioned earlier. In particular, with our approach we were able to achieve the following objectives:

1. Identify Customer Satisfaction in call center conversations. As shown by Rafaeli et al. (2007), this metrics is predicted by so called "customer oriented behaviors";
2. Identify Root Cause of Problems by looking at controversial topics and how agents are able to deal with them;
3. Identify customers who need particular attention based on history of problematic interactions;
4. Learn best practices in dealing with customers by identifying agents able to carry cooperative conversations. This knowledge coupled with customer profiles can be used effectively in intelligent skill-based routing[1]

The article is organized as follows: in section 2 we review current Speech Analytics technology and make the case for Interaction Mining approach in order to address the current business challenges in call centers quality monitoring and assessment. In section 3 we present our Interaction Mining solution based on a specific kind of pragmatic analysis: the argumentative analysis and its implementation with the A3 algorithm. In section, 4 we present the four business cases outlined above, by showing the analysis of actual call center data and the implementation of new relevant metrics and KPIs for call center quality monitoring. We conclude the article with a discussion on the achieved results and a roadmap for future work.

[1] http://en.wikipedia.org/wiki/Skills-based_routing

2 Customer Interaction Analytics Needs Interaction Mining

Call center data contain a wealth of information that usually remains hidden. Key Performance Indicators (KPIs) for call centers performance can be classified into three broad categories (Baird 2004):

1. Agent Performance Statistics: these include metrics such as *Average Speed of Answer*, *Average Hold Time*, *Call Abandonment Rate*, *Attained Service Level*, and *Average Talk Time*. They are based on quantitative measurements that can be obtained directly through ACD[2] Switch Output and Network Usage Data.
2. Peripheral Performance Data: these include metrics such as *Cost Per Call*, *First-Call Resolution Rate*, *Customer Satisfaction*, *Account Retention*, *Staff Turnover*, *Actual vs. Budgeted Costs*, and *Employee Loyalty*. These metrics are mostly quantitative, with the exception of *Customer Satisfaction* that is usually obtained through Customer Surveys. The quantitative metrics are usually collected through Sales Records Expense Records, Human Resources Service Records, and Financial Records.
3. Performance Observation: these include metrics such as *Call Quality*, *Accuracy and Efficiency*, *Adherence to Script*, *Communication Etiquette*, and *Corporate Image Exemplification*. These are qualitative metrics based on analysis of recorded calls and session monitoring by a supervisor.

Minnucci (2004) reports that the most required metrics by call center managers are indeed the qualitative ones topped by Call Quality (100%) and Customer Satisfaction (78%). However, these metrics are difficult to implement with the adequate level of accuracy[3]. For instance, the Baird study (2004) points out that for Customer Satisfaction, accuracy can be:

> "negatively affected by insufficient number of administered surveys per agent resulting in not enough samples of individual agent's work to constitute a representative sample. The result could be an unfair judgment of the agent's performance and allocations of bonuses based more upon chance, good fortune than merit."

This problem would disappear if a more systematic analysis would be conducted over the entire corpus of recorded calls with no human intervention on observation. Therefore, turning such a qualitative metric into a quantitative one is certainly an important challenge that could move Customer Interaction Analytics a leap forward.

Most call center quality monitoring dashboards are now only able to display information related to service-level measures (Agents and Peripheral Performance

[2] Automatic Call Distribution.

[3] Accuracy is defined in (Baird 2004) as true indication and it depends on the actual level of performance attainment, especially with regard to statistical validity.

data), namely how fast and how many calls agents able to handle. Fig. 1 is an example of such a dashboard.

Because of recent improvements of Automatic Speech Recognition (ASR) technology (Neustein 2010), *Speech Analytics* is viewed as a key element implementing call center quality monitoring where ASR technology is leveraged to implement relevant KPIs. As pointed out by Gavalda and Schlueter (2010), Speech Analytics is becoming

> "an indispensable tool to understand what is the driving call volume and what factors are affecting agents' rate of performance in the contact center."

However, current Speech Analytics solutions have focused on *search* rather than *information extraction*. Most system allows the user to search the occurrences of keywords or key-phrases in the spoken conversations (i.e. audio files). While this represents an important feature for targeted call monitoring, it fails in delivering an understanding of the context where such terms occur. In other words, this method can be helpful when the supervisor has a clear idea of what to look for in recorded conversation but it helps very little in data mining and intelligence.

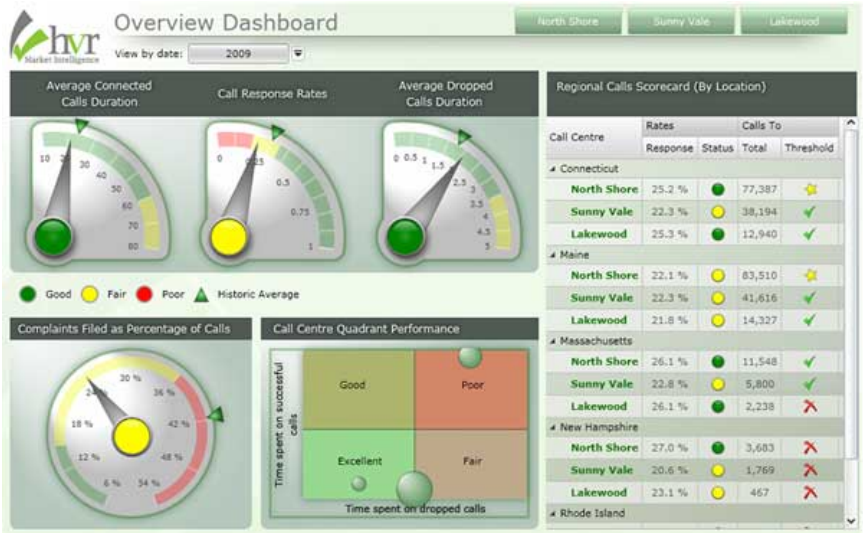

Fig. 1. An example of Call Center dashboard[4] implementing standard metrics

2.1 Interaction Mining

Interaction Mining is an emerging field in Business Analytics that contrasts the standard approach based on Text Mining (Feldman and Sanger 2006). In Text Mining the assumption made is that input is textual and can be treated just as a

[4] An example of Customer Interaction Analytics dashboard is available at http://demos7.dundas.com/HVR.aspx

container of semantic content where non-content words can be filtered out. This assumption is no longer valid in conversational input. Non-content words such as conjunctions, prepositions, personal pronouns and interjections are extremely important in conversations and cannot be filtered out as they bear most of their *pragmatic meaning*. Hence Text Mining tools cannot be entirely transferred to Interaction Mining. Some of the Text Mining tools are still useful such as Entity Recognition, Tokenization and Part-of-Speech Tagging. As pointed out in Pallotta et al. (2011) there are several advantages of moving to Interaction Mining for generating intelligence from conversational content.

It is important to note that while the purpose is similar – i.e., turning unstructured data into structured data for performing quantitative analysis - Text Mining focuses on pattern extraction from *documents*. This is justified because of the massive presence of content words in textual input (e.g. news, articles, blogs, corporate documents). As we mentioned earlier, this is no longer the case with conversational content. The units of information in conversational content are *dialogue turns* and typically they are significantly shorter than documents considered as input for Text Mining. This means that the input has to be linguistically processed in order to understand its *pragmatic function* in the conversation. For instance, a simple turn containing just one single word like "Yes" or "No" can make a substantial difference in the interpretation of a whole conversation. In other words, in Text Mining, the conversational meaning and the context are likely to get lost, which is instead kept and taken into account in Interaction Mining.

Interaction Mining tools are substantially different than those employed in Text Mining. First of all, statistical or extensive machine learning approaches are no longer a viable option since data are very sparse. While it is possible to learn patterns from content-bearing documents, it is nearly impossible to learn pragmatic meaning from non-content bearing words. Approaches that attempted to apply machine learning to Interaction Mining have failed in providing satisfactory results so far (Rienks and Verbree 2006; Hakkani-Tür 2009). This can be explained because the amount of data needed for supervised learning becomes an insurmountable bottleneck. The requirements for Interaction Mining tools are that conversational units, the turns, have to be interpreted in their linguistic context. It is simply not possible to consider them as isolated input, as it is the case with documents or web pages. Moreover, accuracy requirements are higher than those required for Text Mining. For instance, in text categorization, content-words can be used for discrimination and highly frequent non-content words can be removed. This is not the case in conversations where highly frequent non-content words cannot be removed as they carry pragmatic meaning (e.g. prepositions, conjunctions). In Pallotta et al. (2011) we have provided evidences that bag-of-words approach simply is not suitable for pragmatic indexing of conversations, and therefore useless for tasks as Question Answering or Summarization.

Another limitation of Text Mining approach to conversations is in Sentiment Analysis. As Delmonte and Pallotta (2011) previously showed, shallow linguistic processing and machine learning often provide misleading results. Therefore, we

advocated for a deep linguistic understanding of input data even for standalone contributions such as product reviews. In Interaction Mining, the Sentiment Analysis issues become even more compelling because sentiment about a topic is not fully condensed in a single turn but it develops along the whole conversation. For example, it is very common that dissent is expressed toward the opinion of other speakers rather than to the topic under discussion. Sentences like: "why do you think product X is bad?" would be simply mistakenly classified as carrying a negative attitude to product X in a bag-of-word approach.

2.2 Related Work

Current approaches to Customer Interaction Analytics are mostly based on Speech Analytics and Text Mining, which is essentially Search and Sentiment Analysis. Recorded speech is first indexed and searched against a set of negative terms and relevant topics. There are currently two main approaches for speech indexing: i) phonetic transcription and ii) Large Vocabulary Conversational Speech Recognition (LVCSR).

In phonetic transcription, speech recognition is made against a small set of phonemes. Keyword queries are algorithmically converted into their phonetic representation and used for search the phonetic index. With this approach one can search for occurrence of specific terms in calls. Its simplicity is at the same time its strength and weakness. On the one hand the method is fast and accurate but, on the other hand, it is limited to its applicability for generating adequate insights on calls because the context of word's occurrence is lost and it can only recovered by physically listening to the audio excerpt where the searched word occurs. One possible workaround is to systematically search for the occurrence of a large number of terms taken from a pre-defined list, thus obtaining a partially transcribed speech. The words list is typically generated by harvesting domain-related words from corporate documents or from relevant web search results. Several companies put this method forward as the most viable solution for content-based Speech Analytics (see Gavalda and Schlueter (2010) for a detailed coverage).

LVCSR is instead based on the recognition of a very large vocabulary of words and thus provide standard textual transcription of the calls. Transcription errors are usually caused by out-of-vocabulary words and measured as Word Error Rate (WER) [5]. While still considerably high compared to human performance in transcription, accuracy of LVCSR systems show a promising trend as reported by the NIST Speech-To-Text Benchmark Test History 1988-2007 (Fiscus et al. 2008).

Another common approach to the analysis of call center data is that of automatic call categorization through supervised machine learning (Gilman et al. 2004; Zweig et al. 2006; Takeuchi et al. 2009). These methods have failed in providing satisfactory results even for very broad categories. The problem still lies on data sparseness and that huge amount of training data is necessary to achieve reasona-

[5] http://en.wikipedia.org/wiki/Word_error_rate.

ble discriminatory power. Getting huge training data is not an option also because training is highly influenced by domain specificity. Transferring trained models from a domain to another would be problematic.

Unsupervised learning provides better results for domain-specific classes as shown in Tang et al. (2003). However, the sensitivity to the domain represents a big issue. Moreover, this type of categorization – i.e. topics of calls – helps little to understand if a call is satisfactory or not. It might be better suited for retrieval and aggregation of other quality-oriented information.

Instead of downgrading the analysis capabilities we believe it is more appropriate to make the analysis less sensitive to WER and domain-specificity. In other words, we want a robust solution capable of delivering approximate but still sound measurements, which are relevant for implementing call-center metrics. We will show in the next sections that our approach to Interaction Mining is robust and it can properly deal with output from LVCSR systems.

3 Argumentative Analysis for Interaction Mining

Our approach to pragmatic analysis for Interaction Mining is rooted on *argumentative analysis* (Pallotta 2006). Argumentation is a pervasive pragmatic phenomenon in conversations. Purposeful conversations are very often aimed at reaching a consensus for a decision or to negotiate opinions about relevant topics. Both types of conversations contain argumentative actions that can be recognized by Interaction Mining systems. Recognizing the argumentative structure of a conversation is useful for several tasks such as Question Answering, Summarization and Business Analytics (Pallotta et al. 2011). In the specific case of Customer Interaction Analytics, we used argumentative analysis as the basis to perform analysis of conversations and synthesize a cooperativeness score for each of them, which in turn it allows us to predict customer satisfaction for these calls.

In this section we provide a few insights on automatic argumentative analysis performed through the tailoring of an advanced Natural Language Understanding (NLU) technology. Further details are available in Pallotta and Delmonte (2011) and Delmonte et al. (2010).

3.1 Argumentative Structure of Conversations

The argumentative structure defines the various patterns of argumentation used by participants in the dialog, as well as their organization and synchronization in the discussion. A dialog is decomposed into several stages such as issues, proposals, and positions, each stage being possibly related to specific aggregations of elementary dialog acts. Moreover, argumentative interactions may be viewed as specific parts of the discussion where several dialog acts are combined to build such an interaction; for instance, a disagreement could be seen as an aggregation of

several acts of reject and accept of the same proposal. From this perspective, we adopted an argumentative coding scheme, the Meeting Description Schema (MDS), developed in Pallotta (2006). In MDS, the argumentative structure of a meeting is composed of a set of topic discussion episodes, where an episode is a discussion about a specific topic. In each discussing topic, there exists a set of episodes in which several issues are discussed. An issue is generally a local problem in a larger topic to be discussed and solved. Participants propose alternatives, solutions, opinions, ideas, etc. in order to achieve a satisfactory decision. Proposal can be accepted or challenged through acts rejecting or asking questions. Hence, for each issue, there is a corresponding set of proposals episodes (solutions, alternatives, ideas, etc.) that are linked to a certain number of related positions episodes (for example a rejection to a proposed alternative in a discussing issue) or questions and answers.

We illustrate this approach by contrasting the limitation of classical term-based indexing for retrieving relevant content of a conversation. Consider the conversation excerpt in Fig. 2 (a) and the query: "*Why was the proposal on microphones rejected*?". A classical indexing schema would retrieve the first turn from David and by matching the relevant query term "microphone". There is no presence of other query terms such as "reject", "proposal". Moreover, it is not possible to map the "Why" question onto some query term (e.g. reason, motivation, justification, explanation). This makes impossible to adequately answer this query without any additional metadata that highlight the role of the participants' contributions in the conversation.

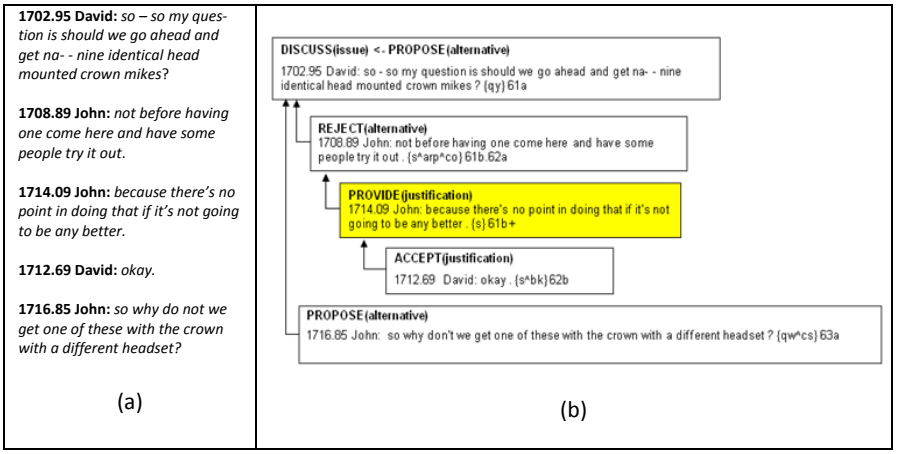

Fig. 2. Argumentative Structure of a conversation

In Fig. 2 (b), we have computed the argumentative structure of the conversation excerpt that allows us to correctly answer the question by selecting the third turn. In fact, the "Why" question is mapped to a query term, which is found as an argumentative index, "justification", for that turn. Of course, finding justification is

not enough, but the retrieval algorithm needs to check whether that justification has been provided as a rejection of a "proposal" (or "alternative") made to an issue on the topic of microphones. This can only be done by navigating back in the argumentative chain up to the turn tagged as "issue" and whose content matches the term "microphone".

3.2 Automatic Argumentative Annotation

The core of our solution is a system that automatically extracts the argumentative structure of conversations, as the one shown in Fig. 2. This system is based on specific tailoring and extension of GETARUNS (Delmonte 2007; 2009), a system for text understanding developed at the University of Venice. *Automatic Argumentative Annotation* (A3) is carried out by a special module of the GETARUNS system activated at the very end of the computation of each dialog. This module takes as input the complete semantic representation produced by the system. To produce Argumentative annotation, the system uses the following 21 Discourse Relations labels: ***statement, narration, adverse, result, cause, motivation, explanation, question, hypothesis, elaboration, permission, inception, circumstance, obligation, evaluation, agreement, contrast, evidence, hypothesis, setting, prohibition.***

These are then mapped onto five general argumentative labels:

1. ACCEPT,
2. REJECT/DISAGREE
3. PROPOSE/SUGGEST
4. EXPLAIN/JUSTIFY
5. REQUEST.

The 5 general argumentative categories are broke down into 12 finer grained categories such as "accept_explanation", "accept_suggestion", "provide_explanation", "reject_explanation", "reject_suggestion", etc. as described in the MDS coding schema.

Details of the A3 algorithm are available in Pallotta and Delmonte (2011) and in Delmonte et al. (2010). The system has been evaluated on conversations from the ICSI meeting corpus (Janin et al. 2001) and annotated by Pallotta et al. (2007). On a total of 2304 turns, 2251 have received an argumentative automatic classification, with a Recall of 97.53%. We computed Precision as the ratio between Correct Argumentative Labels/ Argumentative Labels Found, which corresponds to 81.26%. The F-score is 88.65%.

3.3 Robustness of the A3 algorithm for speech input

The A3 algorithm was evaluated against a corpus of manually transcribed conversations. In order to test if it achieves comparable results on automatically transcribed conversations, we conducted an experiment on similar meetings that have been transcribed using state-of-the-art LVCSR technology (Fiscus et al. 2008). We

have measured the performance of our system and observed an overall degradation of only 11.7% on automatically transcribed conversations with a system showing an average WER of 30% (Hain et al. 2009). This result has been obtained without any intervention on the A3 algorithm itself, which was designed initially to deal with manually transcribed data. These results are quite promising and, coupled with expected improvements in LVCSR technology and further tuning of the system, they provide us with a solid basis for development.

3.4 Multi-word expressions

One key issue with conversation is that topics are not expressed by single words but very often by compounds. Hence, quality of topic detection can be improved if the lexicon contains domain-specific multi-word expressions. We thus run a multi-word expressions extraction tool (Seretan and Wehrli. 2009) to identify the most frequent collocations in the corpus and compare them with the topics detected by the GETARUNS system. The top 10 extracted multi-word expressions[6] (a) and topics (b) are shown in Table 1.

Multi Word Expression	Score
1. Calling Chase	475.4809
2. Account number	300.2746
3. Gross balance	282.5876
4. Direct deposit	247.4588
5. Savings account	189.3173
6. And available	186.6647
7. Social security number	159.8058
8. Area code	146.8807
9. Daytime phone number	143.3286
10. Most recent	126.4333

a

Topic	% of total
1. Chase	5.26%
2. Social security number	3.41%
3. Checking account	2.75%
4. Moment	2.33%
5. Statement	2.16%
6. Money	2.00%
7. Savings account	1.45%
8. Dollar	1.37%
9. Days	1.35%
10. Phone number	1.23%

b

Table 1. Multi Word Expression extracted from the corpus (a) and GETARUN topics (b).

While there is a predictable overlapping it is interesting to see that some domain-dependent terms were detected by the multi-word extraction system but they were not included in the lexicon of our system such as “gross balance” and “and availa-

[6] The score for multi-word expression represents the log-likelihood ratio statistics representing the association strength between the component words (Dunning 1993).

ble", and "direct deposit"[7]. We hypothesize that enriching the lexicon with these terms would greatly improve the pragmatic analysis of conversations. In our future work, we will investigate a way to systematically include multi-word expression information in our system.

4 Business Cases with Call Center Data

In this section we describe how the output of the A3 algorithm can be used to perform Interaction Mining and we present the results of an experiment where we applied it to actual call center data. The main goal was to find out if the argumentative analysis, coupled with other standard text mining analysis (i.e. Sentiment and Subjectivity analysis), could indeed provide useful information to implement several Customer Interaction Analytics metrics. The results show, in particular, that we were able to achieve the objectives we introduced in Section 1. As already discussed in Section 2, being able to extract the above information enables us to implement the relevant and most requested KPIs in call center quality management.

4.1 The Corpus

In our experiment we used a corpus of 213 manually transcribed conversations of a help desk call center in the banking domain. Each conversation has an average of 66 turns and an average of 1.6 calls per agent. This corpus was collected and annotated for a study aimed at identifying conversational behaviors that could favor satisfactory interaction with customers (Rafaeli et al. 2007). This study has shown that is the case and that *customer-oriented behaviors* (COBs) can indeed be used to predict customer ratings. Table 2 contains the identified COBs and their distribution in the annotated portion of the dataset.

Customer Oriented Behaviors	
Anticipating Customers Requests	22,45%
Educating The Customer	16,91%
Offering Emotional Support	21,57%
Offering Explanations / Justifications	28,57%
Personalization Of Information	10,50%

Table 2. Customer-Oriented Behaviors from Call Center data

Notice that only a very small portion of the dataset was manually annotated with COBs, representing only 2.5% of the entire corpus. This prevented us to perform a

[7] The "and" and "Available" are detected as a multi-word expression because they occur frequently in the corpus as the pattern "Gross and Available balance".

statistically sound correlation study, as we cannot consider the non-annotated data as examples of conversations not containing COBs. Additionally, it was not possible to assume that turns that not received COB annotations were actually negative examples, i.e. they did not contain COBs. They were simply not annotated. This prevented us to use the dataset with machine learning algorithms.

4.2 Argumentative analysis of call center conversations

Despite the above limitations, we carried out an experiment by running the A3 algorithm on the COB annotated portion of the dataset. Examples of COB and argumentative annotations of the corpus are shown in Table 3. The table contains turns from a single conversation, which have been annotated as COBs and that are also automatically classified as relevant argumentative categories.

In the example, we show only those turns whose COB and argumentative annotations can be considered as "compatible".

Customer Oriented Behaviors	*Argumentative Categories*		
	Accept explanation	Provide explanation	Suggest
Anticipating Customers Requests			3
Cause if you like I can help you do a transfer by using our touchtone service.			1
Cause with the ATM card I can help you do a transfer by using your touchtone.			1
I can send you a form that you can fill out and return back to us.			1
Offering Emotional Support	1		
Thanks for calling Chase and enjoy your day Miss Asher.	1		
Offering Explanations / Justifications		1	
And that application will be out in the mail within two business days.		1	

Table 3. Examples of COBs from a conversation and their classification as argumentative categories.

We noticed that COBs showed a high resemblance to our argumentative categories and that they might correlate as well. The chart of Fig. 3 shows the number of COB-annotated turns (y-axis) with their argumentative labels assigned by the A3 algorithm, which classifies most of turns containing COBs as "Provide Explanation/Justification" and "Suggest". This does not, of course, prove that there is a statistical correlation between these argumentative categories and COBs because the dataset does not contain negative examples, i.e. turns that are not COB-annotated are not necessarily those that not contain COBs. We have constructed a mapping table, shown in Table 4, which maps argumentative categories onto a numerical scale of cooperativeness. With this mapping, a large proportion of turns that received COB annotation receive positive cooperativeness score.

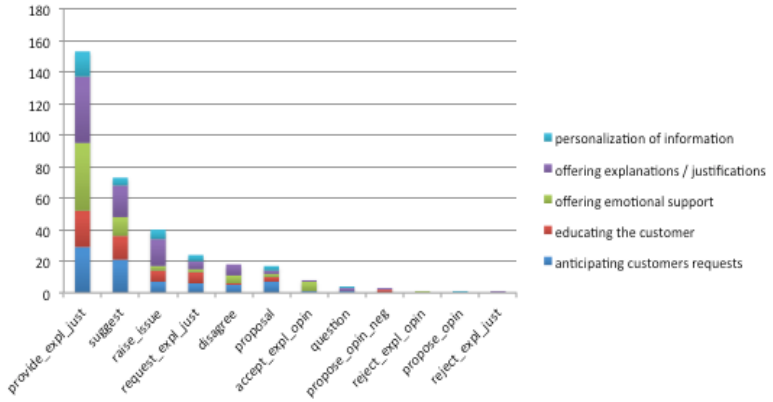

Fig. 3. Correlation between argumentative categories and customer-oriented behaviors.

Argumentative Categories	*Level of Cooperativeness*
Accept explanation	5
Suggest	4
Propose	3
Provide opinion	2
Provide explanation or justification	1
Request explanation or justification	0
Question	-1
Raise issue	-2
Provide negative opinion	-3
Disagree	-4
Reject explanation or justification	-5

Table 4. Mapping table for argumentative categories to levels of cooperativeness

The cooperativeness score is a measure obtained by averaging the score obtained by mapping argumentative labels of each turn in the conversation into a [-5 +5] scale. The mapping is hand crafted and inspired by Bales's Interaction Process Analysis framework (Bales, 1950), where uncooperativeness (i.e. negative scores) is linked to high level of criticism, which is not balanced by constructive contributions (e.g. suggestions and explanations). This mapping provides a reasonable indicator of controversial (i.e. uncooperative) conversations.

In Fig. 4, we can observe how COBs are distributed over cooperativeness score. This provides us with an intuition that cooperativeness score can be a predictor of COBs. We need to stress however that from the dataset we cannot assume that those turns that are not COB annotated do not actually contain COBs.

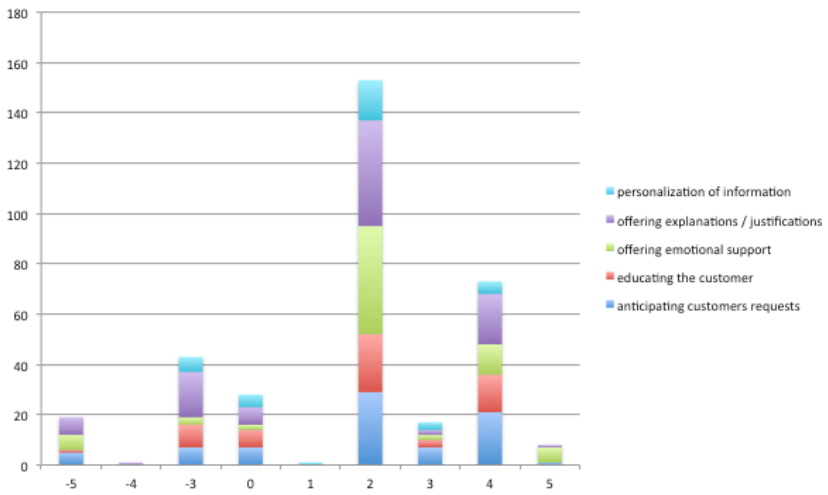

Fig. 4. Distribution of COB annotated turns over cooperativeness scores.

Since the corpus were only partially COB annotated, the only conclusion we can draw from this dataset is that when a turn is COB annotated, there is a high chance that the A3 algorithm would classify it with an argumentative category, which is in turn mapped into a positive cooperativeness score. We cannot conclude instead that turns receiving argumentative annotations that are mapped onto negative cooperativeness scores would actually correspond to absence of COBs.

Our, so far not validated but highly intuitive assumption is that negative cooperativeness scores are also predictors of absence of COBs. This means that turns receiving negative cooperativeness scores can be recognized as elements of customer dissatisfaction. We have reviewed the dataset for those turns that received negative cooperativeness scores and realized that this intuition was reasonably justified. We will provide a more formal assessment of this hypothesis in future work once we have annotated the remaining portion of the dataset.

4.3 Interaction Analytics for Call Center Conversations

We used the Tableau[8] visualization system, which revealed to be a suitable tool for getting insightful multi-dimensional aggregations. We assembled the charts into dashboards that can be used to obtain appropriate summarized information to address the four objectives mentioned in the beginning of this section. We will review each of these objectives and present the related generated dashboard from the analyzed call center data.

Identify Customer Satisfaction

The implementation of the Customer Satisfaction KPI (CSAT) is a direct consequence of the ability to predict COBs. In fact, from Rafaeli et al. (2007), a positive significant correlation exists between customer ratings and COBs.

We pushed this concept a little bit further and we crafted a combined CSAT score by combining cooperativeness score, sentiment and subjectivity analysis. Then turn scores are averaged over the whole conversation. The idea is that customer satisfaction is the result of the overall interaction between the customer and the representative, and not just the occurrence of certain words with positive or negative connotation. We believe that our implementation of CSAT will prove to be more effective and accurate than those based on Sentiment Analysis only.

Combined with additional extracted information such as Sentiment and Subjectivity (see Pallotta and Delmonte (2011) for details), we might safely conclude that CSAT can be predicted by argumentative analysis. Unfortunately, we cannot provide a fully quantitative proof for this claim as the data were not fully annotated and negative examples (i.e. customer dissatisfaction) are missing. As we already said, we consider refining our study towards this direction in future work.

Identify Root Cause of Problems

By looking at controversial topics we can identify root cause of problems in call centers. We selected the worst 20 topics ranked according to frequency of negative attitudes obtained by the Sentiment Analysis module. Fig. 5 shows a dashboard that can be used to detect controversial topics and thus help in spotting unsolved issues.

The user can select one topic and display who addressed that topic (agent or customer) and see the cooperativeness score of each speaker. The rationale in crossing sentiment information with cooperativeness score is a better understanding of the context of sentiment analysis. If, for instance a negative word is used in

[8] http://www.tableausoftware.com

cooperative context (e.g. “providing an explanation”) then its impact in the determination of a cause of problem should be diminished.

Fig. 5. Problem spotting dashboard

For instance, if the representative says: “in order to avoid this problem, you should remove the virus from the computer”, clearly this cannot be considered a negative statement because the representative is simply providing an explanation. This example highlights the limitations of current sentiment-based Speech Analytics solutions, which might be overcome by adopting our approach.

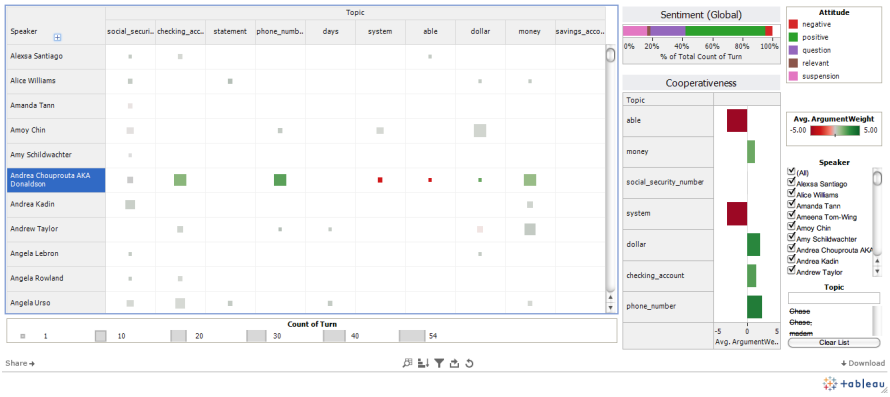

Fig. 6. Topic and Behaviors Dashboard

Another useful tool is that shown in Fig. 6. This dashboard highlights the top 10 most discussed topics and how cooperatively speakers discuss these topics. In the main pane, rows correspond to speakers and for each topic the level of cooperativeness is displayed as a square whose dimension represents the number of turns centered around that topic and the color represents the cooperativeness score. The

cooperativeness histogram shows the overall cooperativeness score for each topic and it is refined when the user selects a specific speaker.

Identify problematic customers

A critical issue in this domain is that customers are not all the same and need to be treated differently according to their style of interaction. There are agents with interpersonal skills who are able to comfortably deal with demanding customers. Agents who show consistently positive cooperativeness can be assumed to be more suitable to deal with extreme cases. Customers who have already shown negative or uncooperative attitudes could be routed to more skilled agents in order to maximize the overall call center performance (i.e. customer satisfaction).

We present a dashboard where problematic customers can be identified and given a particular care. The main tool for this task is a dashboard for speaker assessment shown in Fig. 7. With this dashboard speakers (agents or customers) are ranked according to their cooperativeness score. In the right-hand pane, also the sentiment analysis results are displayed and compared to the overall sentiment score. The analyst can then drill through a specific customer and visualize a specific customer and the calls he/she made.

Fig. 7. Speaker Assessment Dashboard

Once drilled down to a specific call, another useful tool for enabling the detection of problematic customers is the conversation graph (Ailomaa et al. 2008), also integrated in our system. The dashboard shown in Fig. 8 reveals interesting facts about the selected call. In the lower pane, calls are ranked according to their average cooperativeness score. By looking at specific calls, the analyst can display the conversation graph that plots the interaction over the call's timeline (i.e. turns on X-axis). The Y-axis represents the cooperativeness score of each turn. In the right

pane, Sentiment and Argumentative breakdowns are presented for the selected call. Looking closely at calls with conversation graphs helps the supervisor to understand some interaction patterns. If the customer asks for explanations and the representative fails in providing them, the call's cooperativeness score will be lower. The intuition behind this is that challenging turns must be balanced by collaborative turns (e.g. explanations must be provided, suggestions must be given to raised issues).

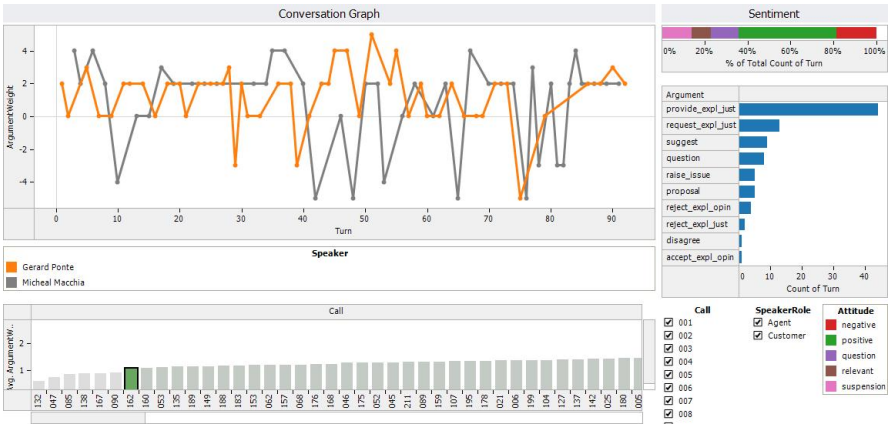

Fig. 8. Conversation graph dashboard

The analyst can then further drill down into the graph or other charts' elements and look at the call's turns as shown in **Error! Reference source not found.**.

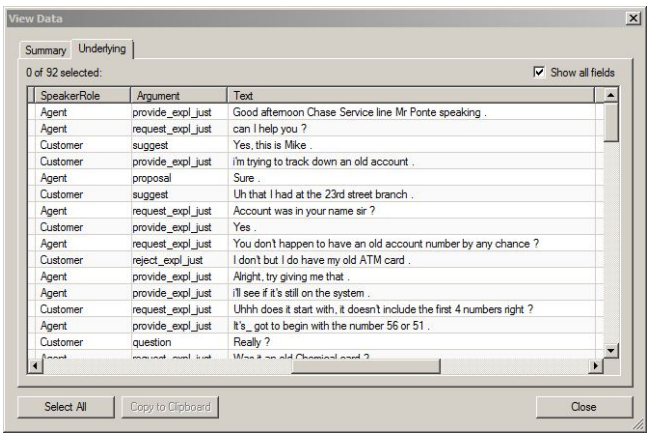

Fig. 9. Drilling through the call's turns

With a different type of representation, our analysis can also serve for exception-based quality management. With the help of a *control chart* displayed in Fig. 10,

we can have an overall look of calls quality (measured by the cooperativeness score) and be warned of the presence of outliers (i.e. calls whose cooperativeness score falls outside the control limits [-3σ, +3σ], which can be in turn analyzed in more detail.

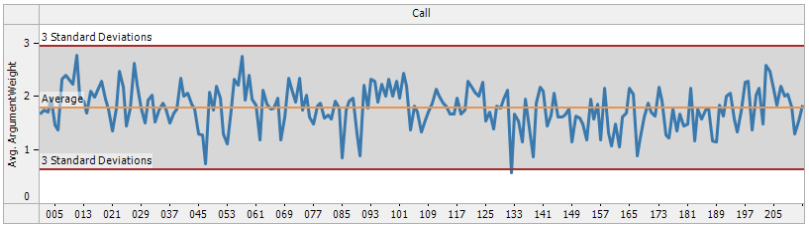

Fig. 10. Control chart of cooperativeness score

Learn best practices from conversations

The implementation of this objective results from considering all the insights gained through the presented visualizations. In particular, Fig. 7 with Agent filtering activated allows one to visualize overall and specific agent's behavior. Best scoring agents can be taken as models and their interaction used as models.

While most of available solutions for skill-based inbound call routing are based on ACD information such as area codes for agent's language selection or based on IVR[9] for option selection. Additionally, the agent selection is often based on efficiency measures in order to optimize the costs and workload (e.g. by assigning the fastest agent to the longest queue). If this strategy might maximize efficiency, they are insufficient to maximize customer satisfaction. We advocate for skill-based call routing based on interpersonal qualities and by influencing the agent selection by cooperativeness requirements.

5 Conclusions

In this article we have presented a new approach to Customer Interaction Analytics based on Interaction Mining, contrasting the current approach based on Speech Analytics and Text Mining. We presented an Interaction Mining tool, which is built on pragmatic analysis of conversations based on argumentation theory. This tool allows us to automatically annotate turns in transcribed conversations with argumentative categories by highlighting the argumentative function of each turn in the conversation. We also showed that our system is robust enough to deal with automatically transcribed speech, as it would be the case in the Call Center do-

[9] Interactive Voice Response

main. An experiment was conducted to see the impact of this technology to a real case. We applied our tool to a corpus of transcribed call center conversations in the banking domains and presented the extracted information in several fashions with the goal of implementing relevant KPIs for Call Center Quality Management.

As for future work we would like to explore other pragmatic dimensions beyond argumentation. This might be relevant in the Call Center domain to look at COBs that are more related to emotional support or providing personalized information, which do not directly relate to argumentation.

We need also to consider finer granularity in argumentative analysis, for instance at clause level. This might be helpful when a single turn carries several argumentative functions. This would definitively improve the quality of the analysis. Our goal is to implement other KPIs for the Call Center domain such as adherence to scripts and corporate image exemplification. In order to achieve these challenging objectives, new types of pragmatic analysis will be required.

Additionally, we would like to explore the possibility of automatically learning agent and customer profiles from our analysis in order to implement more effective skill-based call routing.

Finally, we will annotate the corpus with negative examples, namely turns that do not contain Customer Oriented Behavior. However, the simple absence of COBs does not automatically entail that the turn contains behaviors that could be perceived by the customers as uncooperative. Therefore, a new model will be required where a new class of behavior needs to be identified to model behavior that might lead to customer dissatisfaction. Unfortunately, the corpus does not contain conversations that received low customer ratings. Probably even more challenging will be the task of finding a corpus of call center conversations that contains them.

6 References

Ailomaa M. (2009) Answering Questions About Archived, Annotated meetings. PhD thesis N° 4512, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Baird H. (2004). Ensuring Data Validity Maintaining Service Quality in the Contact Center. Telecom Directions LLC.

Delmonte R. (2007). *Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*, Nova Science Publishers, New York.

Delmonte R. (2009). *Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution*, Nova Science Publishers, New York.

Delmonte R., Bristot A. and Pallotta V. (2010). Deep Linguistic Processing with GETARUNS for spoken dialogue understanding. In *Proceedings of LREC 2010 conference*. Malta 18-23 May, 2010.

Delmonte R. and Pallotta V. (2011). Opinion Mining and Sentiment Analysis Need Text Understanding. In Pallotta V., Soro A., Vargiu E. (eds.). *Advances in Distributed Agent-Based Retrieval Tools*. Series: Studies in Computational Intelligence, Springer Verlag.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence , Computational Linguistics, 19(1): 61-74.

Feldman, R., and Sanger, J. (2006). The Text Mining handbook. Advanced approaches in analyzing unstructured data. Cambridge University Press.

Fiscus J.G., Ajot J. and Garofolo J.S. (2008). The Rich Transcription 2007 Meeting Recognition Evaluation. In Stiefelhagen R., Bowers R. and Fiscus J. (Eds.) Multimodal Technologies for Perception of Humans. Lecture Notes In Computer Science, Vol. 4625. Springer-Verlag, Berlin, Heidelberg. pp. 373-389.

Hakkani-Tür D. (2009). Towards automatic argument diagramming of multiparty meetings. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Taipei, Taiwan. April, 2009.

Gavalda M. and Schlueter J. (2010). The Truth is Out There: Using Advanced Speech Analytics to Learn Why Customers Call Help-line Desks and How Effectively They Are Being Served by the Call Center Agent. Chapter 10 of Neustein A. (ed.) *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*. Springer Media, LLC 2010.

Gilman A., Narayanan B. and Paul S. (2004). Mining call center dialog data. In Zanasi A., Ebecken N.F.F. and Brebbia C.A. (Eds.) *Data Mining V*. WIT Press.

Hain T., Burget L., Dines J., Garner P.N., El Hannani A., Huijbregts M., Karafiat M., Lincoln M. and Wan V. (2009). The AMIDA 2009 Meeting Transcription System In *Proceedings of INTERSPEECH 2009*, Makuhari, Japan, pp. 358-361.

Janin A., Baron D., Edwards J., Ellis D., Gelbart D., Morgan N., Peskin B., Pfau T., Shriberg E., Stolcke A. and Wooters C. (2003). The ICSI Meeting Corpus. In *Proceedings of IEEE/ICASSP 2003*, 6-10 April 2003, Hong Kong, vol. 1, pp. 364-367.

Minnucci J. (2004). Call Center KPIs: A Look at How Companies Are Measuring Performance, a special report published by ICMI Inc., 2004, p. 2.

Neustein A. (ed.) (2010). *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, DOI 10.1007/978-1-4419-5951-5_10, Springer Media, LLC 2010.

Pallotta V. (2006), Framing Arguments. In *Proceedings of the International Conference on Argumentation* ISSA, June 2006, Amsterdam, Netherlands.

Pallotta V. and Delmonte R. (2011), Argumentative Models for Interaction Mining. Journal of Argument and Computation. 2(2), 2011.

Pallotta V., Seretan V. and Ailomaa M. (2007). User requirements analysis for Meeting Information Retrieval based on query elicitation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (ACL 2007), pages 1008–1015, Prague.

Pallotta V., Vrieling L. and Delmonte R. (2011). Interaction Mining: Making business sense of customers conversations through semantic and pragmatic analysis. In Zorrilla M., Mazón J-N., Ferrández Ó., Garrigós I., Daniel F. and Trujillo J. (Eds.) *Business Intelligence Applications and the Web: Models, Systems and Technologies*. IGI Global to appear.

Rafaeli, A., Ziklik, L., & Doucet, L. (2007). The impact of call center employees' customer orientation behaviors on customer satisfaction. *Journal of Service Research*.

Rienks R. and Verbree D. (2006). About the usefulness and learnability of argument-diagrams from real discussions. In *Proceedings of the 3rd International Machine Learning for Multimodal Interaction Workshop* (MLMI 2006), May 1-4, 2006, Bethesda (MD), USA.

Seretan, Violeta and Eric Wehrli (2009). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, 43(1), 71–85.

Tang M., Pellom, B. and Hacioglu, K. (2003). Call-type classification and unsupervised training for the call center domain. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding ASRU '03*, pp. 204 – 208.

Takeuchi H., Subramaniam L.V., Nasukawa T. and Roy S. (2009). Getting insights from the voices of customers: Conversation mining at a contact center. *Information Sciences* n. 179 pp. 1584–1591, Elsevier.

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 *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume*, pages 292–295, New York City, Association for Computational Linguistics.