Automatic Identification of Important Segments and Expressions for Mining of Business-Oriented Conversations at Contact Centers

Hironori Takeuchi[*], **L Venkata Subramaniam**[†], **Tetsuya Nasukawa**[*], **and Shourya Roy**[†]

[*]IBM Research, Tokyo Research Laboratory
Shimotsuruma 1623-14, Yamato-shi
Kanagawa 2428502 Japan
{hironori, nasukawa}@jp.ibm.com

[†]IBM Research, India Research Laboratory
Institutional Area 4, Block-C, Vasant Kunj
New Delhi 110070 India
{lvsubram, rshourya}@in.ibm.com

Abstract

Textual records of business-oriented conversations between customers and agents need to be analyzed properly to acquire useful business insights that improve productivity. For such an analysis, it is critical to identify appropriate textual segments and expressions to focus on, especially when the textual data consists of complete transcripts, which are often lengthy and redundant. In this paper, we propose a method to identify important segments from the conversations by looking for changes in the accuracy of a categorizer designed to separate different business outcomes. We extract effective expressions from the important segments to define various viewpoints. In text mining a viewpoint defines the important associations between key entities and it is crucial that the correct viewpoints are identified. We show the effectiveness of the method by using real datasets from a car rental service center.

1 Introduction

"Contact center" is a general term for customer service centers, help desks, and information phone lines. Many companies operate contact centers to sell their products, handle customer issues, and address product-related and services-related issues. In contact centers, analysts try to get insights for improving business processes from stored customer contact data. Gigabytes of customer contact records are produced every day in the form of audio recordings of speech, transcripts, call summaries, email, etc. Though analysis by experts results in insights that are very deep and useful, such analysis usually covers only a very small (1-2%) fraction of the total call volume and yet requires significant workload. The demands for extracting trends and knowledge from the whole text data collection by using text mining technology, therefore, are increasing rapidly.

In order to acquire valuable knowledge through text mining, it is generally critical to identify important expressions to be monitored and compared within the textual data. For example, given a large collection of contact records at the contact center of a manufacturer, the analysis of expressions for products and expressions for problems often leads to business value by identifying specific problems in a specific product. If 30% of the contact records with expressions for a specific product such as "ABC" contain expressions about a specific trouble such as "cracked", while the expressions about the same trouble appear in only 5% of the contact records for similar products, then it should be a clue that the product "ABC" may actually have a crack-related problem. An effective way to facilitate this type of analysis is to register important expressions in a lexicon such as "ABC" and "cracked" as associated respectively with their categories such as "product" and "problem" so that the behavior of terms in the same category can be compared easily. It is actually one of the most important steps of text mining to identify such relevant expressions and their categories that can potentially lead to some valuable insights. A failure in this step often leads to a failure in the text mining. Also, it has been considered an artistic task that requires highly experienced consul-

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 458–467, Prague, June 2007.

tants to define such categories, which are often described as the viewpoint for doing the analysis, and their corresponding expressions through trial and error.

In this paper, we propose a method to identify important segments of textual data for analysis from full transcripts of conversations. Compared to the written summary of a conversation, a transcription of an entire conversation tends to be quite lengthy and contains various forms of redundancy. Many of the terms appearing in the conversation are not relevant for specific analysis. For example, the terms for greeting such as "Hello" and "Welcome to (Company A)" are unlikely to be associated with specific business results such as purchased-or-not and satisfied-or-not, especially because the conversation is transcribed without preserving the nonverbal moods such as tone of voice, emotion etc. Thus it is crucial to identify key segments and notable expressions within conversations for analysis to acquire valuable insights.

We exploit the fact that business conversations follow set patterns such as an opening followed by a request and the confirmation of details followed by a closing, etc. By taking advantage of this feature of business conversations, we have developed a method to identify key segments and the notable expressions within conversations that tend to discriminate between the business results. Such key segments, the trigger segments, and the notable expressions associated with certain business results lead us to easily understand appropriate viewpoints for analysis.

Application of our method for analyzing nearly one thousand conversations from a rental car reservation office enabled us to acquire novel insights for improving agent productivity and resulted in an actual increase in revenues.

Organization of the Paper: We start by describing the properties of the conversation data used in this paper. Section 3 describes the method for identifying useful viewpoints and expressions that meet the specified purpose. Section 4 provides the results using conversational data. After the discussion in Section 5, we conclude the paper in Section 6.

2 Business-Oriented Conversation Data

We consider business-oriented conversation data collected at contact centers handling inbound telephone sales and reservations. Such business oriented conversations have the following properties.

- Each conversation is a one-to-one interaction between a customer and an agent.
- For many contact center processes the conversation flow is well defined in advance.
- There are a fixed number of outcomes and each conversation has one of these outcomes.

For example, in car rentals, the following conversation flow is pre-defined for the agent. In practice most calls to a car rental center follow this call flow.

- Opening - contains greeting, brand name, name of agent
- Pick-up and return details - agent asks location, dates and times of pick up and return, etc.
- Offering car and rate - agent offers a car specifying rate and mentions applicable special offers.
- Personal details - agent asks for customer's information such as name, address, etc.
- Confirm specifications - agent recaps reservation information such as name, location, etc.
- Mandatory enquiries - agent verifies clean driving record, valid license, etc.
- Closing - agent gives confirmation number and thanks the customer for calling.

In these conversations the participants speak in turns and the segments can be clearly identified. Figure 1 shows part of a transcribed call.

Each call has a specific outcome. For example, each car rental transaction has one of two call types, reservation or unbooked, as an outcome.

Because the call process is pre-defined, the conversations look similar in spite of having different results. In such a situation, finding the differences in the conversations that have effects on the outcomes is very important, but it is very expensive and difficult to find such unknown differences by human analysis. We show that it is possible to define proper viewpoints and corresponding expressions leading to insights on how to change the outcomes of the calls.

```
AGENT: Welcome to CarCompanyA. My name is Albert.  How may I
help you?
.........
AGENT: Allright may i know the location you want to pick the
car from.
CUSTOMER: Aah ok I need it from SFO.
AGENT: For what date and time.
.........
AGENT: Wonderful so let me see ok mam so we have a 12 or 15
passenger van avilable on this location on those dates and
for that your estimated total for those three dates just
300.58$ this is with Taxes with surcharges and with free
unlimited free milleage.
.........
AGENT : alright mam let me recap the dates you want to pick
it up from SFO on 3rd August and drop it off on august 6th in
LA alright
CUSTOMER : and one more questions Is it just in states or
could you travel out of states
.........
AGENT : The confirmation number for your booking is 221 384.
CUSTOMER : ok ok Thank you
Agent :  Thank you for calling CarCompanyA and you have a
great day good bye
```

Figure 1: Transcript of a car rental dialog (partial)

3 Trigger Segment Detection and Effective Expression Extraction

In this section, we describe a method for automatically identifying valuable segments and concepts from the data for the user-specified difference analysis. First, we present a model to represent the conversational data. After that we introduce a method to detect the segments where the useful concepts for the analysis appear. Finally, we select useful expressions in each detected trigger segment.

3.1 Data Model

Each conversational data record in the collection D is defined as $\boldsymbol{d}_i$. Each $\boldsymbol{d}_i$ can be seen as a sequence of conversational *turns* in the conversational data, and then $\boldsymbol{d}_i$ can be divided as

$$\boldsymbol{d}_i = \boldsymbol{d}_i^1 + \boldsymbol{d}_i^2 + \cdots + \boldsymbol{d}_i^{M_i}, \qquad (1)$$

where $\boldsymbol{d}_i^k$ is the k-th turn in $\boldsymbol{d}_i$ and M_i is the total number of turns in $\boldsymbol{d}_i$. The $+$ operator in the above equation can be seen as an equivalent of the string concatenation operator. We define $\boldsymbol{d}_i^{\sim j}$ as the portion of $\boldsymbol{d}_i$ from the beginning to turn j. Using the same notation, $\boldsymbol{d}_i^{\sim j} = \boldsymbol{d}_i^1 + \boldsymbol{d}_i^2 + \cdots + \boldsymbol{d}_i^j$. The collection of $\boldsymbol{d}_i^{\sim m_k}$ constitutes the *Chronologically Cumulative Data* up to turn m_k (D_k). D_k is represented as

$$D_k = (\boldsymbol{d}_1^{\sim m_k}, \boldsymbol{d}_2^{\sim m_k}, \ldots, \boldsymbol{d}_n^{\sim m_k}). \qquad (2)$$

Figure 2 shows an image of the data model. We set some m_k and prepare the chronologically cumulative data set as shown in Figure 3. We represent binary mutually exclusive business outcomes such as success and failure resulting from the conversations as “A” and “not A”.

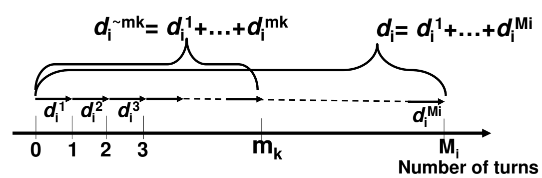

Figure 2: Conversation data model

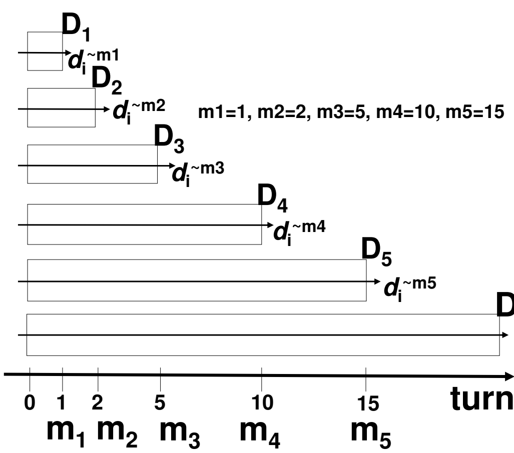

Figure 3: Chronologically cumulative conversational data

3.2 Trigger Segment Detection

Trigger segments can be viewed as portions of the data which have important features which distinguish data of class “A” from data of class “not A”.

To detect such segments, we divide each chronologically cumulative data set D_k into two data sets, training data $D_k^{training}$ and test data D_k^{test}. Starting from D_1, for each D_k we trained a classifier using $D_k^{training}$ and evaluated it on D_k^{test}. Using accuracy, the fraction of correctly classified documents, as a metric of performance (Yang and Liu, 1999), we denote the evaluation result of the categorization as $acc(categorizer(D_k))$ for each D_k and plot it along with its turn. Figure 4 shows the effect of gradually increasing the training data for the classification. The distribution of expressions

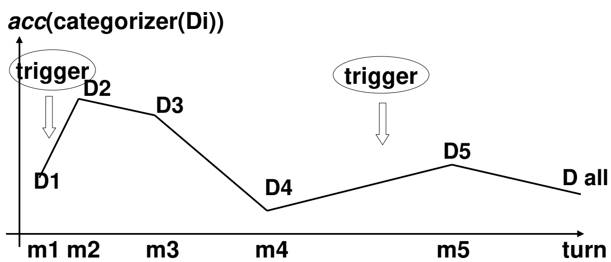

Figure 4: Plot of $acc(categorizer(D_k))$

in a business-oriented conversation will change almost synchronously because the call flow is predefined. Therefore $acc(categorizer(D_k))$ will increase if features that contribute to the categorization appear in D_k. In contrast, $acc(categorizer(D_k))$ will decrease if no features that contribute to the categorization are in D_k. Therefore, from the transitions of $acc(categorizer(D_k))$, we can identify the segments with increases as triggers where the features that have an effect on the outcome appear. We denote a trigger segment as $seg(start\,position, end\,position)$. Because the total numbers of turns can be different, we do not detect the last section as a trigger. In Figure 4, $seg(m_1, m_2)$ and $seg(m_4, m_5)$ are triggers. It is important to note that using the cumulative data is key to the detection of trigger segments. Using non-cumulative segment data would give us the categorization accuracy for the features within that segment but would not tell us whether the features of this segment are improving the accuracy or decreasing it. It is this gradient information between segments that is key to identifying trigger segments.

Many approaches have been proposed for document classification (Yang and Liu, 1999). In this research, however, we are not interested in the classification accuracy itself but in the increase and decrease of the accuracy within particular segments. For example, the greeting, or the particular method of payment may not affect the outcome, but the mention of a specific feature of the product may have an effect on the outcome. Therefore in our research we are interested in identifying the particular portion of the call where this product feature is mentioned, along with its mention, which has an effect on the outcome of the call. In our experiments we used the SVM (Support Vector Machine) classifier (Joachims, 1998), but almost any classifier should work because our approach does not depend on the classification method.

3.3 Effective Expression Extraction

In this section, we describe our method to extract effective expressions from the detected trigger segments.

The effective expressions in D_k are those which are representative in the selected documents and appear for the first time in the trigger segments $seg(m_i, m_j)$. Numerous methods to select features exist (Hisamitsu and Niwa, 2002) (Yang and Pedersen, 1997). We use the χ^2 statistic for each expression in D_k as a representative metric. For the two-by-two contingency table of a expression w and a class "A" shown in Table 1, the χ^2 statistic is calculated as

Table 1: Contingency table for calculating the χ^2 statistic

	# of documents including w	# of documents not including w
A	n_{11}	n_{12}
not-A	n_{21}	n_{22}

$$\chi^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})} \quad (3)$$

where N is the number of documents. This statistic can be compared to the χ^2 distribution with one degree of freedom to judge representativeness.

We also want to extract the expressions that have not had an effect on the outcome before D_k. To detect the new expressions in D_k, we define the metric

$$new(w) = \frac{w(D_k)}{\max(w(D_{k-1}),1)} / \frac{m_k}{m_{k-1}} \times \mathrm{sign}(w(D_k^A) - w(D_k^{notA})), \quad (4)$$

where $w(D_k)$ is the frequency of expression w in the chronologically cumulative data D_k, $\max(a, b)$ selects the larger value in the arguments, m_k is the number of turns in D_k, $w(D_k^A)$ is the frequency of w in D_k with the outcome of the corresponding data being "A", and $\mathrm{sign}(\cdot)$ is the signum function. When w in class "A" appears in D_k much more frequently than D_{k-1} compared with the ratio of their turns, this metric will be more than 1. We detect significant expressions by considering the combined score $\chi^2(w) \cdot new(w)$. Using this combined score, we can filter out the representative expressions that have already appeared before D_k and distinguish significant expressions that first appear in D_k for each class "A" and "not A".

3.4 Appropriate Viewpoint Selection

In a text mining system, to get an association that leads to a useful insight, we have to define appropriate viewpoints. Viewpoints refer to objects in relation to other objects. In analysis using a conventional text mining system (Nasukawa and Nagano, 2001), the viewpoints are selected based on expressions in user dictionaries prepared by domain experts. We have identified important segments of the conversations by seeing changes in the accuracy of a categorizer designed to segregate different business outcomes. We have also been able to extract effective expressions from these *trigger* segments to define various viewpoints. Hence, viewpoint selection is now based on the trigger segments and effective expressions identified automatically based on specified business outcomes. In the next section we apply our technique to a real life dataset and show that we can successfully select useful viewpoints.

4 Experiments and Results

4.1 Experiment Data and System

We collected 914 recorded calls from the car rental help desk and manually transcribed them. Figure 1 shows part of a call that has been transcribed.

There are three types of calls:

1. **Reservation Calls:** Calls which got *converted*. Here, "converted" means the customer made a reservation for a car. Reserved cars can get *picked-up* or *not picked-up*, so some reserved cars do not eventually get picked-up by customers (no shows and cancellations).

2. **Unbooked Calls:** Calls which did not get converted.

3. **Service Calls:** Customers changing or enquiring about a previous booking.

The distribution of the calls is given in Table 2.

Table 2: Distribution of calls

Unbooked Calls	461
Reservation Calls (Picked-Up)	72
Reservation Calls (Not Picked-Up)	65
Service Calls	326
Total Calls	914

The reservation calls are most important in this context, so we focus on those 137 calls. In the reservation calls, there are two types of outcomes, car picked-up and car not picked-up. All reservation calls look similar in spite of having different outcomes (in terms of pick up). The reservation happens during the call but the pick up happens at a later date. If we can find differences in the conversation that affect the outcome, it is expected that we could improve the agent productivity. Reservation calls follow the pre-defined reservation call flow that we mentioned in Section 2 and it is very difficult to find differences between them manually. In this experiment, by using the proposed method, we try to extract trigger segments and expressions to find viewpoints that affect the outcome of the reservation calls.

For the analysis, we constructed a text mining system for the difference analysis "picked-up" vs. "not picked-up". The experimental system consists of two parts, an information extraction part and a text mining part. In the information extraction part we define dictionaries and templates to identify useful expressions. In the text mining part we define appropriate viewpoints based on the identified expressions to get useful associations leading to useful insights.

4.2 Results of Trigger Segment Detection and Effective Expression Extraction

Based on the pre-defined conversation flow described in Section 2, we set m_1=1, m_2=2, m_3=5, m_4=10, m_5=15, and m_6=20 and prepared $D_1, \ldots, D_6$ and D. The features of $\boldsymbol{d}_i$ consist of nouns, compound nouns, specified noun phrases (e.g. adjective+noun), and verbs. For each D_k we calculated $acc(categorizer(D_k))$ for the classes "picked-up" and "not picked-up." In this process, we use a SVM-based document categorizer (Joachims, 2002). Of the 137 calls, we used 100 calls for training the categorizer and 37 calls for trigger segment detection. Figure 5 shows the results of $acc(categorizer(D_k))$ for picked-up. The accuracy of classification using the data of entire conversations ($acc(categorizer(D))$ is 67.6% but we are trying to detect important segments by considering not the accuracy values themselves but the gradients between segments. From these results, $seg(1, 2)$ and

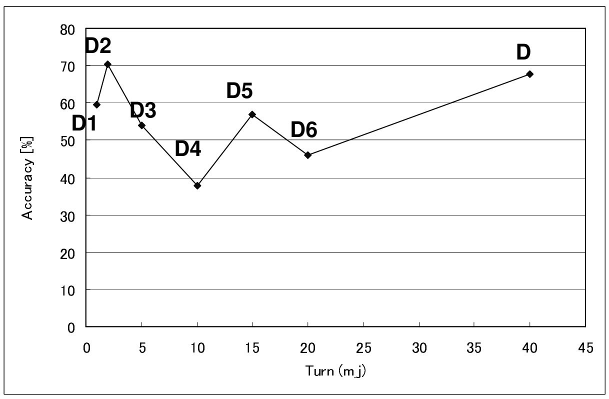

Figure 5: Result of $acc(categorizer(D_k))$

$seg(10, 15)$ are detected as trigger segments. We now know that these segments are highly correlated to the outcome of the call.

For each detected trigger segment, we extract effective expressions in each class using the metric described in Section 3.3. Table 3 shows some expressions with high values for the metric for each trigger. In this table, "just NUMERIC dollars" is a canonical expression and an expression such as "just 160 dollars" is mapped to this canonical expression in the information extraction process. From this result, in $seg(1, 2)$, "make", "reservation" are correlated with "pick up" and "rate" and "check" are correlated with

Table 3: Selected expressions in trigger segments

Trigger	Selected expressions	
	pick up	not picked up
$seg(1, 2)$	make, return, tomorrow, day, airport, look, assist, reservation, tonight	rate, check, see want, week
$seg(10, 15)$	number, corporate program, contract, card, have, tax surcharge, just NUMERIC dollars, discount, customer club, good rate, economy	go, impala

"not-picked up". By looking at some documents containing these expressions, we found customer intention phrases such as "would like to make a reservation", "want to check a rate", etc. Therefore, it can be induced that the way a customer starts the call may have an impact on the outcome. From expressions in $seg(10, 15)$, it can be said that discount-related phrases and mentions of the good rates by the agent can have an effect on the outcome.

We can directly apply the conventional methods for representative feature selection to D. The following expressions were selected as the top 20 expressions from whole conversational data by using the χ^2 metric defined in (3).

corporate program, contract, counter, September, mile, rate, economy, last name, valid driving license,BRAND NAME, driving, telephone, midsize, tonight, use, credit, moment, airline, afternoon

From these results, we see that looking at the call as a whole does not point us to the fact that discount-related phrases, or the first customers-utterance, affect the outcome. Detecting trigger segments and extracting important expressions from each trigger segment are key to identifying subtle differences between very similar looking calls that have entirely opposite outcomes.

4.3 Results of Text Mining Analysis using Selected Viewpoints and Expressions

From the detected segments and expressions we determined that the customer's first utterance along with discount phrases and value selling phrases affected the call outcomes. Under these hypotheses, we prepared the following semantic categories.

- Customer intention at start of call: From the customer's first utterance, we extract the following intentions based on the patterns.
  - strong start: *would like to make a booking*, *need to pick up a car*, ...
  - weak start: *would like to check the rates*, *want to know the rate for vans*, ...

  Under our hypotheses, the customer with a strong start has the intention of booking a car and we classify such a customer as a **booking_customer**. The customer with a weak start usually just wants to know the rates and is classified as a **rates_customer**.
- discount-related phrases: *discount, corporate program, motor club, buying club* ... are registered into the domain dictionary as discount-related phrases.
- value selling phrases: we extract phrases mentioning good rates and good vehicles by matching patterns related to such utterances.
  - mentions of good rates: *good rate*, *wonderful price*, *save money*, *just need to pay this low amount*, ...
  - mentions of good vehicles: *good car*, *fantastic car*, *latest model*, ...

Using these three categories, we tried to find insights to improve agent productivity.

Table 4 shows the result of two-dimensional association analysis for 137 reservation calls. This table shows the association between customer types based on customer intention at the start of a call and pick up information. From these results, 67%

Table 4: Association between customer types and pick up information

Customer types extracted from texts based on customer intent at start of call	Pick up information	
	pick up	not-picked up
booking_customer (w/ strong start) (70)	47	23
rates_customer (w/ weak start) (37)	13	24

(47 out of 70) of the booking_customers picked up the reserved car and only 35% (13 out of 37) of the rates_customers picked it up. This supports our hypothesis and means that pick up is predictable from the customer's first or second utterance.

It was found that cars booked by rates_customers tend to be "not picked up," so if we can find any actions by agents that convert such customers into "pick up," then the revenue will improve. In the booking_customer case, to keep the "pick up" high, we need to determine specific agent actions that concretize the customer's intent.

Table 5 shows how mentioning discount-related phrases affects the pick up ratios for rates_customers and booking_customers. From this table, it can

Table 5: Association between mention of discount phrases and pick up information

Rates_customer	Pick up information	
Mention of discount phrases by agents	pick up	not-picked up
yes (21)	10	11
no (16)	3	13
Booking_customer	Pick up information	
Mention of discount phrases by agents	pick up	not picked up
yes (40)	30	10
no (30)	17	13

be seen that mentioning discount phrases affects the final status of both types of customers. In the rates_customer case, the probability that the booked car will be picked up, $P(\text{pick-up})$ is improved to 0.476 by mentioning discount phrases. This means customers are attracted by offering discounts and this changes their intention from "just checking rate" to "make a reservation here". We found similar trends for the association between mention of value selling phrases and pick up information.

4.4 Improving Agent Productivity

From the results of the text mining analysis experiment, we derived the following actionable insights:

- There are two types of customers in reservation calls.
  - **Booking_customer** (with strong start) tends to pick up the reserved car.
  - **Rates_customer** (with weak start) tends not to pick up the reserved car.
- In the **rates_customer** case, "pick up" is improved by mentioning discount phrases.

By implementing the actionable insights derived from the analysis in an actual car rental process, we verified improvements in pick up. We divided the 83 agents in the car rental reservation center into two groups. One of them, consisting of 22 agents, was trained based on the insights from the text mining analysis. The remaining 61 agents were not told about these findings. By comparing these two groups over a period of one month we hoped to see how the actionable insights contributed to improving agent performance. As the evaluation metric, we used the pick up ratio - that is the ratio of the number of "pick-ups" to the number of reservations.

Following the training the pick up ratio of the trained agents increased by 4.75%. The average pick up ratio for the remaining agents increased by 2.08%. Before training the ratios of both groups were comparable. The seasonal trends in this industry mean that depending on the month the bookings and pickups may go up or down. We believe this is why the average pick up ratio for the remaining agents also increased. Considering this, it can be estimated that by implementing the actionable insights the pick up ratio for the pilot group was improved by about 2.67%. We confirmed that this difference is meaningful because the p-value of the t-test statistic is 0.0675 and this probability is close to the standard t-test (α=0.05). Seeing this, the contact center trained all of its agents based on the insights from the text mining analysis.

5 Discussion

There has been a lot of work on specific tools for analyzing the conversational data collected at contact centers. These include call type classification for the purpose of categorizing calls (Tang et al., 2003) (Zweig et al., 2006), call routing (Kuo and Lee, 2003) (Haffner et al., 2003), obtaining call log summaries (Douglas et al., 2005), agent assisting and monitoring (Mishne et al., 2005), and building of domain models (Roy and Subramaniam, 2006). Filtering problematic dialogs automatically from an automatic speech recognizer has also been studied (Hastie et al., 2002) (Walker et al., 2002). In contrast to these technologies, in this paper we consider the task of trying to find insights from a collection of complete conversations. In (Nasukawa and Nagano, 2001), such an analysis was attempted for agent-entered call summaries of customer contacts by extracting phrases based on domain-expert-specified viewpoints. In our work we have shown that even for conversational data, which is more complex, we could identify proper viewpoints and prepare expressions for each viewpoint. Call summaries by agents tend to mask the customers' intention at the start of the call. We get more valuable insights from the text mining analysis of conversational data. For such an analysis of conversational data, our proposed method has an important role. With our method, we find the important segments in the data for doing analyses. Also our analyses are closely linked to the desired outcomes.

In trigger detection, we created a chronologically cumulative data set based on turns. We can also use the segment information such as the "opening" and "enquiries" described in Section 2. We prepared data with segment information manually assigned, made the chronologically cumulative data and applied our trigger detection method. Figure 6 shows the results of $acc(categorizer(D_k))$. The trend in

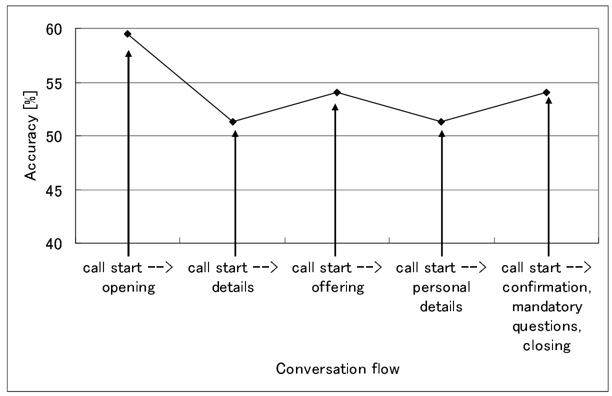

Figure 6: Result of $acc(categorizer(D_k))$ using segment information

Figure 6 is similar to that in Figure 5. From this result, it is observed that "opening" and "offering" segments are trigger segments. Usually, segmentation is not done in advance and to assign such information automatically we need data with labeled segmentation information. The results show that even in the absence of labeled data our trigger detection method identifies the trigger segments. In the experiments in Section 4, we set turns for each chronologically cumulative data by taking into account the pre-defined call flow.

In Figure 5 we observe that the accuracy of the categorizer is decreasing even when using increasing parts of the call. Even the accuracy using the complete call is less than using only the first turn. This indicates that the first turn is very informative, but it also indicates that the *features* are not being used judiciously. In a conventional classification task, the number of features are sometimes restricted when constructing a categorizer. It is known that selecting only significant features improves the classification accuracy (Yang and Pedersen, 1997). We used *Information Gain* for selecting features from the document collection. This method selects the most discriminative features between two classes. As expected the classification accuracy improved significantly as we reduced the total number of features from over 2,000 to the range of 100 to 300. Figure 7 shows the changes in accuracy. In the pro-

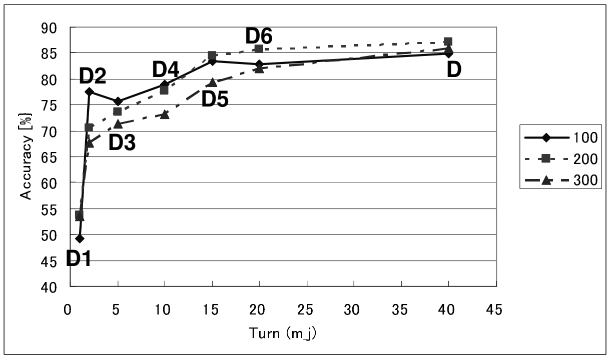

Figure 7: Result of $acc(categorizer(D_k))$ with top 100 to 300 features selected using information gain

posed method, we detect trigger segments using the increases and decreases of the classification accuracy. By selecting features, the noisy features are not added in the segments. Decreasing portions, therefore are not observed. In this situation, as a trigger segment, we can detect the portion where the gradient of the accuracy curve increases. Also using feature selection, we find that the classification accuracy is highest when using the entire document, which is expected. However, we notice that the trigger segments obtained with and without feature selection are almost the same.

In the experiment, we use manually transcribed data. As future work we would like to use the noisy output of an automatic speech recognition system to obtain viewpoints and expressions.

6 Conclusion

In this paper, we have proposed methods for identifying appropriate segments and expressions automatically from the data for user specified difference analysis. We detected the trigger segments using the property that a business-oriented conversation follows a pre-defined flow. After that, we identified the appropriate expressions from each trigger segment. It was found that in a long business-priented conversation there are important segments affecting the outcomes that can not been easily detected by just looking through the conversation, but such segments can be detected by monitoring the changes of the categorization accuracy. For the trigger segment detection, we do not use semantic segment information but only the positional segment information based on the conversational turns. Because our method does not rely on the semantic information in the data, therefore our method can be seen as robust. Through experiments with real conversational data, using identified segments and expressions we were able to define appropriate viewpoints and concepts leading to insights for improving the car rental business process.

Acknowledgment

The authors would like to thank Sreeram Balakrishnan, Raghuram Krishnapuram, Hideo Watanabe, and Koichi Takeda at IBM Research for their support. The authors also appreciate the efforts of Jatin Joy Giri at IBM India in providing domain knowledge about the car rental process and thank him for help in constructing the dictionaries.

References

S. Douglas, D. Agarwal, T. Alonso, R. M. Bell, M. Gilbert, D. F. Swayne, and C. Volinsky. 2005. Mining customer care dialogs for "daily news". *IEEE Transaction on Speech and Audio Processing*, 13(5):652–660.

P. Haffner, G. Tur, and J. H. Wright. 2003. Optimizing svms for complex call classification. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 632–635.

H. W. Hastie, R. Prasad, and M. A. Walker. 2002. What's the trouble: Automatically identifying problematic dialogues in darpa communicator dialogue systems. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 384–391.

T. Hisamitsu and Y. Niwa. 2002. A measure of term representativeness based on the number of co-occurring sailent words. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pages 137–142.

T. Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142.

H.-K J. Kuo and C.-H. Lee. 2003. Discriminative training of natural language call routers. *IEEE Transaction on Speech and Audio Processing*, 11(1):24–35.

G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer. 2005. Automatic analysis of call-center conversations. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 453–459.

T. Nasukawa and T. Nagano. 2001. Text analysis and knowledge mining system. *IBM Systems Journal*, pages 967–984.

S. Roy and L. V. Subramaniam. 2006. Automatic generation of domain models for call centers from noisy transcriptions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING/ACL)*, pages 737–744.

M. Tang, B. Pellom, and K. Hacioglu. 2003. Call-type classification and unsupervised training for the call center domain. In *Proceesings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 204–208.

M. A. Walker, I. Langkilde-Geary, H. W. Hastie, J. Wright, and A. Gorin. 2002. Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, 16:393–319.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49.

Y. Yang and J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*, pages 412–420.

G. Zweig, O. Shiohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury. 2006. Automatic analysis of call-center conversations. In *Proceedings of IEEE Internatinal Conference of Acoustics, Speech and Signal Processing (ICASSP)*, pages 589–592.