**RC24785 (W0904-078) April 17, 2009**
**Computer Science**

# IBM Research Report

## Low-Cost Call Type Classification for Contact Center Calls Using Partial Transcripts

**Youngja Park, Wilfried Teiken, Stephen C. Gates**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

**Research Division**
**Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich**

# Low-Cost Call Type Classification for Contact Center Calls Using Partial Transcripts

*Youngja Park, Wilfried Teiken, Stephen C. Gates*

IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598, USA
{young_park, wteiken, scgates}@us.ibm.com

## Abstract

Call type classification and topic classification for contact center calls using automatically generated transcripts is not yet widely available mainly due to the high cost and low accuracy of call-center grade automatic speech transcription. To address these challenges, we examine if using only partial conversations yields accuracy comparable to using the entire customer-agent conversations. We exploit two interesting characteristics of call center calls. First, contact center calls are highly scripted following prescribed steps, and the customers problem or request (i.e., the determinant of the call type) is typically stated in the beginning of a call. Thus, using only the beginning of calls may be sufficient to determine the call type. Second, agents often more clearly repeat or rephrase what customers said, thus it may be sufficient to process only agents' speech.

Our experiments with 1,677 customer calls show that two partial transcripts comprising only the agents utterances and the first 40 speaker turns actually produce slightly higher classification accuracy than a transcript set comprising the entire conversations. In addition, using partial conversations can significantly reduce the cost for speech transcription.

**Index Terms**: Call Type Classification, Contact Center Call Analysis, Speech Analytics, Machine Learning

## 1. Introduction

In most contact centers, agents manually categorize calls into a predefined set of call types after handling a new customer call. The call type information is then widely used by contact center management to help understand trends in call volumes, performance of agents on particular types of calls, and other key contact center measures. An automatic system which can transcribe contact center conversations and determine the call type is therefore desirable. Automatic call type classification for telephone conversations, however, is not yet widely available due to the high cost and low accuracy of automatic speech transcription; transcription accuracy of contact center calls typically is in the range of 50% to 75% in our experience.

Contact center calls, however, contain two interesting characteristics that can be used to mitigate the problems. First, contact center calls are usually highly scripted (i.e., they follow prescribed steps), and in those scripts the customer's problem or request (i.e., an important determinant of the call type) is typically described in the beginning of a call. For instance, customer calls to an automotive company's contact center typically begins with the agent's greeting and introducing himself to the customer followed by the customer's description on the problem or request. Second, key portions of the customer's narrative are typically replicated or summarized in the agent's utterances because the script calls for the agent to repeat or rephrase what the customer said to make sure that he understood the customer's requests or concerns. In the contact center environment, automatic speech recognition (ASR) systems tend to show lower error rate for the agent speech because the ASR system is more adapted to the agents, and agents produce higher-quality sound than do customers.

Based on these observations on contact center calls, we hypothesize that we may be able to identify the call type by using only a beginning part or only the agent's utterances of a call and yet still achieve acceptable classification accuracy. Such a system would be desirable because it could significantly reduce the costs involving call recording and speech transcription, and thus enable contact centers to benefit from automatic call classification. To investigate the hypotheses, we examine the effect of using different parts of calls on call type classification. In particular, we conduct experiments with the first $n$ speaker turns (or utterances) for a relatively small number $n$, and the utterances spoken by only one speaker.

In this work, we apply a support vector machine (SVM) for classifying automatically transcribed contact center calls. Support vector machines have been successfully used in previous work on topic categorization [1, 2, 3]. We train the classifier with the call types assigned by the contact center agents for eliminating the need for expensive manual labeling.

The experimental results show that a transcript set comprising only the agents' utterances produce higher classification accuracy and recall than the entire transcripts by 1% and 3% respectively. In addition, the first 40 speakers turns yielded almost same classification accuracy as when the entire conversation was analyzed. The results indicate that transcribing only the agent's channel or the first 40 turns would be sufficient to build a call type classification, resulting in 40% to 64% cost saving.

## 2. Related Work

Automatic topic classification of texts has been advanced significantly in recent years and is being used in many information retrieval and knowledge management applications [1, 4, 5, 6, 2, 3]. Automatic call type classification for contact center calls, however, is not yet widely available due to the high costs of automatic speech transcription, and the poor quality of the automatic transcripts.

A few attempts have recently been made for automatic topic classification of contact center data. Busemann *et al.* presented an automatic topic classification system for e-mail messages re-

ceived in a contact center [7]. Haffner *et al.* applied SVMs for call classification on the spoken language understanding (SLU) component of the the AT&T's *How May I Help You* natural dialog system [8, 9]. This is not, however, designed to process human-to-human conversations in free format. Tang *et al.* described a call-type classification system for an Information Technology (IT) Help Desk call center [10]. Their approach, however, has restrictions that significantly reduce the applicability of the system. First, they constructed the taxonomy by hand after inspecting the data used in the study. Second, the system was trained with human generated call transcripts which are very expensive.

# 3. Call Type Classification System

This section describes the target call types and the classification system used in this work in more detail. Our approach is domain-independent and can be applied to contact centers across different industries. The development of the system is, however, guided by sample customer calls to an automotive company's contact center located in the US.

## 3.1. Classification Method

In this work, we apply support vector machines for call type classification as implemented in LIBSVM [11]. Particularly, we use C-support vector classification (C-SVC) with a radial basis function (RBF) kernel [12]. C-SVC solves the following problem:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$
$$subject\ to\ \ y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, \ldots, l$$

given training vectors $\mathbf{x}_i \in R^n, i = 1, \ldots, l$ and an answer vector $\mathbf{y} \in R^l$.

C-SVC is designed for two-class classification problems. For multi-class ($k$) classification like the problem in this work, LIBSVM uses the "one-against-one" approach in which $\frac{k(k-1)}{2}$ classifiers are constructed, and each one trains data from two different classes.

## 3.2. Target Call Types

The contact center has 61 different call types, and the distribution of service requests across the call types is highly unbalanced having 45.7% of all service requests categorized into a single call type. In this work, we selected the six most frequently used call types, which together constitute 91% of all service requests in the contact center. The target call types include "Complaint Vehicle", "Dealer Issue", "Request for Information (RFI) on Dealer Location"," RFI on Promotions", "RFI on Vehicle" and "RFI on Warranty".

The six call types can be broadly grouped into two categories; complaint calls (i.e. "Complaint Vehicle", "Dealer Issue"), and information-seeking calls (i.e., "RFI" call types). Note that the distinction between "Complaint Vehicle" and "RFI on Vehicle" is vague and rather ad-hoc. Calls in both types discuss various issues related to the customer's vehicle such as an engine problem or a navigation system failure. If, however, a customer is focused on expressing his or her dissatisfaction with the car, the call is categorized into "Complaint Vehicle". If the customer is more interested in seeking a solution, agents are required to categorize such calls into "RFI on Vehicle".

## 3.3. Features

In this work, we exploit two types of lexical features for categorizing calls into the six call types described in Section 3.2. The first group of features consists of word unigrams that appear in at least five different calls (i.e., a bag of word approach). The second group of features includes semantic lexical features which can indicate the speaker's emotion and dissatisfaction. These features are designed to distinguish complaint calls from RFI calls.

**Word features:** Automatic transcripts are ill-written, containing many disfluencies (e.g., "uh" and "umm"), spoken numbers (e.g., "two thousand and three"), spellings (e.g., "j o n e s") and acronyms (e.g., "a b s"). It is important to note that many spoken spellings and acronyms are domain terms that are good indicators of the call type.

In this work, we first conduct *text normalization* to improve the quality of call transcripts to extract more accurate word features. The normalization process includes filler normalization, spelled-out expression and acronym normalization, and word aggregation. We identified 25 different fillers from sample transcripts generated by the ASR system used in this work (more details are described in Section 4.1), and the information about which particular filler is used is not important for call type classification. We thus convert all filler words into an artificial word, $FILLER$, and treat all instances of fillers as a single word. Similarly, all numeric expressions such as telephone numbers and street addresses are converted into an artificial word, $NUM$. Spoken spellings and acronyms are merged into words. For instance, "j o n e s" becomes "jones", and "a b s" becomes "abs". We then perform word aggregation in which all inflectional variants of a word (e.g., "seatbelt" and "seatbelts") are merged into a single canonical form. Finally, we extract words from the normalized transcripts that appear at least five different calls as features. The feature value of a word is the count of the word's occurrences in a call.

**Brand names:** Mentions of brand names include automobile brand or model names manufactured by the competitors as well as by the company. We observed that complaint calls tend to contain product names or a competitor's name more often than RFI calls. For instance, a customer of a "Complaint Vehicle" call might say "I will buy an *XXX* [1] next time".

In this work, we use a manually compiled lexicon containing most major automotive company names and their product names to recognize brand and competitor mentions. Note that, unlike word features, many automobile brand or model names are multiple words. We count the number of brand names mentioned by the agent and by the customer separately.

**Sentiment words:** Complaint calls such as in "Complaint Vehicle" type and in "Dealer Issue" type often contain many sentiment words expressing the customer's emotional status and dissatisfaction. The sentiment words, especially negative sentiment words, are useful to distinguish complaint calls from RFI calls.

To identify words with negative sentiment polarity, we use the subjectivity lexicon described in [13, 14]. The lexicon contains a list of words with a priori prior polarity (*positive*, *negative*, *neutral* and *both*) and the strength of the polarity (*strongsubj* vs. *weaksubj*). In this work, we use only words of which

[1] a competitor's name

prior polarity is either *positive* or *negative*, and the strength of the polarity is *strongsubj*.

We first recognize all strong positive and strong negative words in a call transcript, and then perform a local context analysis to decide the polarity of a sentiment word in the given context. If a sentiment word has a polarity shifter within a two-word window to the left, the polarity of the word is changed based on the shifter [15]. For instance, if a positive sentiment word appears with a negation word such as *no* and *not*, the polarity of the word in the context becomes negative. Like the brand name feature, we count the number of negative sentiment words mentioned by the agent and by the customer separately to identify the holder of the sentiment.

## 4. Experiments

The main question we ask in this work is whether we can identify the call type by analyzing a partial transcript such as the first $n$ speaker turns for a relatively small number $n$, or the utterances spoken by the agent. Our hypotheses behind the question are two-fold. First, the call type can be determined by analyzing only the first $n$ utterances because the customer's question is typically stated in the beginning of a call. Second, the agent's utterances alone would be sufficient to determine the call type because the agent often reveals the customer's question by repeating or summarizing it and by providing the solution to the problem.

To investigate these hypotheses, we conduct experiments with several different sets of partial transcripts and compare the results with the performance obtained using the entire calls. More specifically, we conduct experiments with only the agents' utterances and only the customer's utterances to study the effect of the speaker on call type classification. To find out what is a good value for $n$, we conduct experiments with the transcript sets comprising the first 10, 20, 30 and 40 speaker turns respectively. Hereafter, we call the transcript sets *EntireCall*, *AgentOnly*, *CustomerOnly*, *First10*, *First20*, *First30* and *First40* respectively.

### 4.1. Data

We acquired recorded customer calls to a contact center of the automotive company which were recorded during a two month period time in 2007. We used the IBM Research Attila Speech recognition toolkit to transcribe the contact center calls [16]. The general language model was trained on data from various sources including conversational telephony speech and broadcast news. The acoustic model and the domain-specific language model were trained with 340 hours of customer calls to the contact center in addition to approximately 2,000 hours of general conversational telephony speech data. The final ASR system shows an overall word error rate of 26.4% for the contact center calls.

In addition to the call recordings, we obtained the agent-assigned call types for the recorded calls, which are used as the ground truth for training and performance evaluation. It is important to note that a call type is associated with a service request which often involves more than one call. However, the call type is usually determined when the first call is handled. We, therefore, extracted the first call of the service requests which were assigned to one of the six call types by the agents, resulting in 1,677 calls. The calls have 99.5 speaker turns on average, and the average call duration is 856 seconds.

The distribution of the 1,677 calls across the six agent-assigned call types and detailed size information of the experimental data sets are shown in Table 1 and Table 2 respectively.

| Call Type | Number of Calls | Percentage |
|---|---|---|
| Complaint Vehicle | 380 | 22.7% |
| Dealer Issue | 251 | 15.0% |
| RFI Dealer Location | 160 | 9.5% |
| RFI Promotions | 118 | 7.0% |
| RFI Vehicle | 448 | 26.7% |
| RFI Warranty | 320 | 19.1% |

Table 1: Distribution of the 1,677 calls across the six call types

| Transcript Set | Total Speech Time | Total Number of Words |
|---|---|---|
| *EntireCall* | 322 hours 39 minutes | 2,650,701 |
| *AgentOnly* | 191 hours | 1,361,764 |
| *CustomerOnly* | 131 hours 39 minutes | 1,288,937 |
| *First10* | 28 hours 42 minutes | 290,650 |
| *First20* | 57 hours 25 minutes | 549,206 |
| *First30* | 88 hours | 787,918 |
| *First40* | 118 hours 13 minutes | 1,017,692 |

Table 2: Detailed size information of the experimental data sets

### 4.2. Performance Comparison

In this section, we discuss the experimental results of the automatic call type classification system. All performance measures are computed using a 10-fold cross validation.

Figure 1 shows overall classification accuracy, mean recall, and mean precision of call type classification carried out with *EntireCall*, *AgentOnly* and *CustomerOnly* transcript sets. Mean recall and mean precision are macro-averaged recall and precision across the six call types.

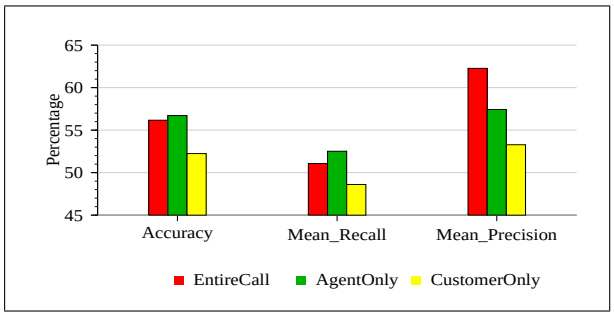


Figure 1: Comparison of classification accuracy, mean precision and mean recall resulted by *EntireCall*, *AgentOnly* and *CustomerOnly* transcript sets.

As we can see from the figure, the highest mean precision was achieved when the entire calls were used. However, the transcript sets comprising only agents' utterances produced higher classification accuracy and recall than the entire transcripts. It is important to note that the performance obtained by using only the customers' utterances is significantly lower than the other two transcript sets. We argue that the reasons behind

the better performance by the *AgentOnly* set are two-fold. First, the automatic speech recognition (ASR) system's error rate is lower for the agent speech because the ASR system is more adapted to the agents, and agents usually produce higher-quality sound than do customers. Second, the agents utterances contain more topical words than the customers' utterances because agents usually both repeat the customer's questions and provide the solutions to the customers using many topical words.

The second experiment is designed to examine whether we can identify the call type by processing only the beginning part of a call. We conduct experiments with the first $n$ speaker turns, with $n$ ranging from 10 to 40 with interval 10. The values were determined based on our analysis on some sample calls, in which the customers' questions or requests are typically described within the range.

Figure 2 depicts how the classification accuracy changes as the number of speaker turns used for classification increases. The circle symbols in Figure 2 show the overall classification accuracy levels when the utterances spoken by both the agent and the customer were used. The triangle symbols show the classification accuracy levels when only the agent's turns extracted from the transcripts sets were used. The line in the figure indicates the accuracy resulted by using *EntireCall*.

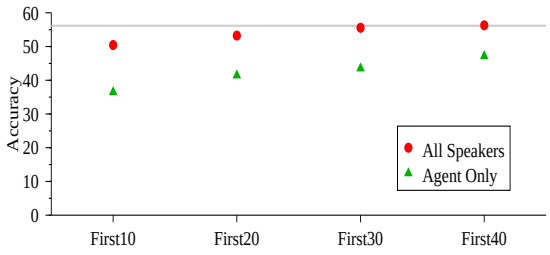


Figure 2: Changes of the classification accuracy as the number of turns increases.

As we can see from the figure, *First30* achieves almost same level of accuracy as *EntireCall*, and *First40* slightly outperforms *EntireCall*. Unlike the result of the first experiment, the agents turns alone don't produce comparable results as when both agent and customer turns were included. The reason for the lower accuracy is that the agents' rephrasing of the customers' requests usually appear in a later part of the calls. In many cases, the agents collect information about the customer such as address and telephone number, and the vehicle such as model and year during the problem description step.

## 5. Conclusions

In this paper, we investigated whether partial conversations of contact center calls can yield a classification accuracy comparable to that of entire conversations based on the following two observations on contact center calls. First, the customer's problem or request, which often determines the call type, is typically stated in the beginning of a call. Second, agent's utterances tend to have better recognition and to contain more topical words, which are very important to identify the call type.

The experimental results suggest that we can achieve slightly superior call type classification system by using only the agent's utterances or the first 40 turns in a call, which provide cost saving by 40% and 64% respectively. [2]

## 6. References


[1] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp. 137 – 142.

[2] E. Leopold and J. Kindermann, "Text categorization with support vector machines. how to represent texts in input space?" *Machine Learning*, vol. 46, pp. 423–444, 2002.

[3] F. Peng, D. Schuurmans, and S. Wang, "Language and task independent text categorization with simple language models," in *Proceedings of HLT-NAACL*, 2003, pp. 110–117.

[4] W. Lam and C. Y. Ho, "Using a generalized instance set for automatic text categorization," in *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 81–89.

[5] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 42–49.

[6] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.

[7] S. Busemann, S. Schmeier, and R. G. Arens, "Message classification in the call center," in *Proceedings of the sixth conference on Applied natural language processing*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 158–165.

[8] P. Haffner, G. Tur, and J. Wright, "Optimizing svms for complex call classification," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, 2003.

[9] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?" *Speech Communication*, vol. 23, pp. 113–127, 1997.

[10] M. Tang, B. Pellom, and K. Hacioglu, "Call-type classification and unsupervised training for the call center domain," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2003.

[11] C. Chang and C. Lin, "Libsvm: a library for support vector machines," Software available at *http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[12] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[13] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "OpinionFinder: A system for subjectivity analysis," in *Proc. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005) Companion Volume (software demonstration)*, 2005.

[14] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada, 2005, pp. 347–354.

[15] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*, 2004.

[16] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 205–208.


[2] The cost reduction rate is computed based on the reduced speech time of the partial transcript sets compared to the entire transcripts.