CALL-TYPE CLASSIFICATION AND UNSUPERVISED TRAINING FOR THE CALL CENTER DOMAIN

Min Tang, Bryan Pellom, Kadri Hacioglu

Center for Spoken Language Research
University of Colorado at Boulder
Boulder, Colorado 80309-0594, USA
{tangm,pellom,hacioglu}@cslr.Colorado.EDU

ABSTRACT

In this paper we describe recent experiments in call-type classification and acoustic modeling for speech recognition in the call center domain. We first describe the CU Call Center Corpus, a database of human-to-human conversations recorded from an Information Technology (IT) Help Desk call center located on the University of Colorado campus. Next, we describe our analysis and labeling of the recorded conversations into a hierarchical taxonomy of the call types. We consider four methods for call-type classification and provide initial experiments illustrating classification error rates for this new task domain. It is shown that lightly supervised training based on using the output from an automatic speech recognizer in conjunction with supervised labeling of calls by call-type can substantially reduce classification error rates and development efforts when only limited training data are available. A call-type classification error rate of 24% is achieved using a classifier based on Support Vector machines. Finally, we consider issues related to unsupervised acoustic and language model training for improved call transcription and point to directions for future work.

1. INTRODUCTION

Customer contact centers, or call centers, represent a new domain for integrating advanced speech technologies for improved customer services. Most notably there has been the trend for older complicated touch-tone menus to be converted into voice-enabled directed dialogs. In recent years several commercial systems have been developed to automatically route calls to particular service agents based on a brief spoken problem description [1, 2, 3]. Call routing systems save money and improve customer satisfaction by connecting the customer to the most appropriate and skilled agent. The router reduces queue time and call duration – saving both time and money while simultaneously improving customer satisfaction. To-date most commercial systems have considered deployments in large-scale call center environments where work forces tend to consist of highly specialized agents that perform specific business tasks. Collaborative group interactions between contact agents within large call centers tends to be low.

Smaller call centers are more abundant and diversified in terms of group interactions needed to solve tasks. In such conditions, agents are more likely to collaborate to address new or unseen problems and share skill sets. We point out the need for new speech and language technologies for call centers that can facilitate collaborative work environments. We believe that careful integration of speech and language systems can help to foster new and effective forms of collaboration between agents while improving customer service, call efficiency, reducing agent workload and training time. To begin to focus on these new directions for call center speech research, this paper presents initial results towards collection and analysis of the University of Colorado Call Center Corpus. We first describe the corpus collection, transcription, and call-type analysis. Calls are labeled into a hierarchical call-type taxonomy. We then consider four machine learning methods for automatically classifying incoming calls based on this call-type labeling. Results are shown both for classifiers trained from human-transcribed data as well as for classifiers trained in a lightly-supervised manner.

Deploying new speech recognition technologies into call centers is expensive and time-consuming due to the cost of transcribing and labeling data for each new task domain. In this paper, we also consider issues related to utilizing untranscribed data as well as HTML text-documents related to the task-domain for improving acoustic and language modeling for call center transcription. Earlier work by Lamel et. al. [4, 5] have shown that given enough untranscribed data and a good language model, the unsupervised training can achieve similar performance with regards to supervised training. In this work we explore this methodology and consider both the issues related to unsupervised acoustic and language model training from the vast amount of data that can be collected from a call center.

2. CU CALL CENTER CORPUS

The CU Call Center Corpus consists of conversations collected from the University of Colorado Information Technology Services (ITS) telephone Help Desk, also known as the IT Service Center. The IT Service Center is a small contact center consisting of approximately 20 agents who support the trouble-shooting of a wide range of computer and telecommunications issues. The call center is staffed daily by three to five agents who typically field 200 calls per day with an average call duration of 5 minutes.

The data collection was conducted in collaboration with Avaya Inc., along with the Center for Spoken Language Research (CSLR), the Department of Communication, and the Alliance for Technology Learning and Society (ATLAS) on the CU Boulder campus. Specialized hardware from Avaya was provided to perform dual-channel audio recording of the telephone conversations. Callers who dialed the contact center during the data collection period were played a message prompt explaining the study. Callers were to allowed to opt-out from participating in the data collection.

The data collection considered calls recorded from late February 2002 through early April 2002. Table 1 summarizes the collected corpus. In total, 4359 conversation sides were collected.

The data totals 283 hours of audio of which 1816 conversation sides have been transcribed at the word level. The transcribed data has been scrubbed for personal information (e.g., last names of callers, passwords, and identifying information such as government issued ID numbers). Regions of audio containing personal information were marked by hand and later replaced with 500 Hz audible tones. Similar procedures were used to remove the identifying personal information in text from the word-level transcriptions. During the data collection period, researchers in the Department of Communication also conducted on-site observations of call center agents to obtain a better understanding call center work-flow, group dynamics, and social interactions amongst call center agents [11].

Table 1. *CU Call Center Corpus*

	Transcribed	Untranscribed	Total
Call Sides	1816	2543	4359
Audio-time	119 hours	164 hours	283 hours
Talk-time	51 hours	51 hours	102 hours

3. CALL-TYPE LABELING

We have examined the transcribed conversations between callers and agents and have developed a hierarchical taxonomy of call-types for this task domain. Our taxonomy is based on analysis of 743 dual-channel transcribed conversations (1486 call-sides). We began our analysis by providing a summary of each call and later grouped similar calls into levels representing broad classes of call types. A total of 98 detailed call types have been determined through inspection of the data. A subset of the taxonomy of calls is shown in Figure 1.

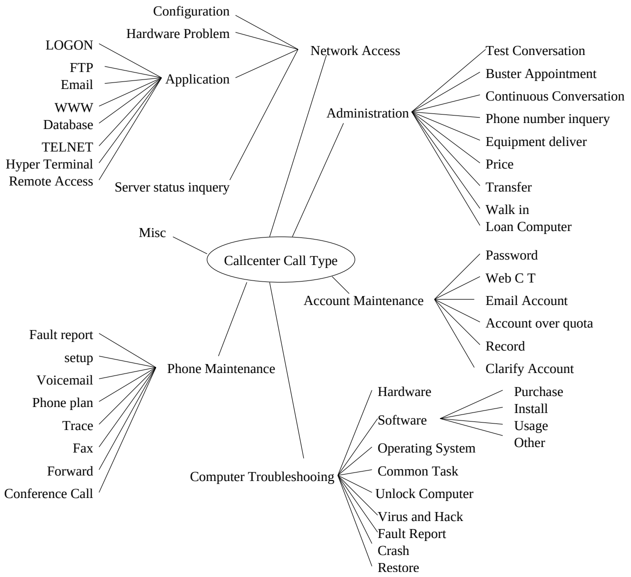

Fig. 1. The hierarchical taxonomy of call types.

We can see from Figure 1 that all calls within the corpus can be described by 6 broad classes – Account Maintenance, Administration, Network Access, Phone Maintenance and Computer trouble shooting, and Miscellaneous. The broad classes are further divided into children and grandchildren nodes (note that grandchildren nodes are not shown in Figure 1). Conversations can be assigned to more than one class label. Table 2 shows the call type distribution for 743 labeled conversations in terms of the 6 broad categories of call types.

Table 2. *Call-type distribution.*

Call-Type	Percentage of Calls
Network Access	38.5 %
Account Maintenance	18.9 %
Administration	12.7 %
Computer Troubleshooting	12.2 %
Phone Maintenance	12.0 %
Miscellaneous	5.7 %

4. CALL-TYPE CLASSIFICATION

Automatic classification of conversations has several call center applications. Such data can provide call center managers with summaries related to percentage of calls fielded within broad categorizations of daily business tasks. It can also be used to pinpoint areas where call efficiency is not well maintained by agents. In this section we consider four different approaches for automatically classifying calls:

Language Model Approach: Calls related to different classes are assumed to differ in terms of word choice and sequence. We can construct a statistical *n*-gram language model (LM) for each class from transcribed conversations. A classifier can be designed by submitting the test conversation to each LM and measuring the output likelihood of the transcription given the class-conditioned language model weighted by the class-prior probability. The classifier chooses the class label with the highest likelihood. Because of data sparsity, we use a back-off bigram language model with Witten-Bell discounting for this work.

Naive Bayes: Naive Bayes is a simple classification algorithm. It assumes that each term in a document (i.e., a conversation in this work) is drawn independently from a multinomial distribution and classifies according to the Bayes optimal decision rule. Each class has its own set of multinomial parameters.

$$C^* = \underset{i}{argmin} \sum_{t_i \in T} -logP(t_i|C_i)P(C_i) \qquad (1)$$

where C^* is the best class, C_i is the class i, t_i is one of the terms in test conversation and T is the test conversation. Terms in this work are represented by word-pairs. We use *BOW lib* ("Bag of words library") developed by Andrew Mccallum to construct the Naive Bayes classifier [8].

TFIDF: The TFIDF algorithm is similar to Naive Bayes except that it uses term-frequency inverse document frequency instead of term-frequency alone. *BOW lib* is also used to construct the TFIDF classifier.

Support Vector Machine: The Support Vector Machine is originally proposed by Vapnik. It finds a maximal margin separating hyperplane between two classes of data. We use the *BOW lib* to pre-process the texts into feature-vector form, and then the open source SVMfu 3.1 software to construct the SVM multiclass classifier [7, 8]. The feature-vector is extracted in this way: after excluding 419 function words, the counts of the top 2500 unigrams and bigrams with highest information gain in the training set are used as the features. Support Vector Machines have gained much attention recently due to their ability to provide superior classification performance for sparse and high-dimensional data problems.

5. UNSUPERVISED TRAINING

In addition to call-type classification, we considered an initial investigation of methods for rapidly porting our existing baseline Switchboard recognition system to a new call center task domain using unsupervised training from untranscribed data.

5.1. Baseline Call Transcription System

Experiments conducted in this paper use Sonic, the University of Colorado large vocabulary continuous speech recognition system [10]. Our baseline transcription system has been trained using 283 hours of telephone data from the Switchboard task. Our trigram language model was developed by SRI for the 2000 NIST Hub5 evaluation and consists of a 34k word vocabulary. The transcription system uses iterative multiple regression class MLLR (mean and variance) adaptation as well as vocal tract length normalization (VTLN). The baseline system has a 32.4% word error rate on the 2000 NIST Switchboard evaluation set.

5.2. Training Method

Our baseline recognition system was used to provide rough transcriptions for the 51 hours of speech activity contained in the untranscribed portion of the CU Call Center Corpus. The transcriptions obtained from the recognizer contain word, phone, and HMM state alignments in addition to word-posterior probabilities. From the 51 hours of untranscribed speech data, we selected a subset of 28 hours of material for experimentation (note that some portions of the data contain acoustic recording issues such as peak-clipping which were discarded using an automated procedure).

Acoustic Modeling: We first consider combining the transcribed Switchboard training data along with the additional automatically transcribed data and incorporate both materials for acoustic training. It is expected that the transcription error rate for conversational telephone speech will be high. To mitigate this problem we have also investigated weighting observations during training by their associated word-posterior probability as output by the recognizer [9]. Transcribed data from Switchboard is assumed to have posterior probability of unity during retraining.

Language Modeling: For language model training using automatically transcribed data, we have utilized confidence information at the utterance level. Similar to the work of Hakkani-Tűr in [6], we obtain utterance level confidence measure by computing the arithmetic mean of word confidence scores across the utterance. We then filter sentences with low confidence and train the new language model based on the remaining utterances. We have also considered incorporation of text data extracted from webpages from the IT Service Center website to improve the language model in a semi-automated fashion. The additional text data from web pages consists of 415k words of material. We combine the high-confidence utterance from the speech recognizer transcription together with the web page data and train a new task-specific *n*-gram language model.

6. EXPERIMENTAL EVALUATION

6.1. Call Type Classification

Experiments have been performed using the 743 hand-labeled conversations from the transcribed portion of the call center corpus. Results are shown in Table 3 for the case of a classifier trained on 668 randomly selected conversations and tested on the remaining 75 conversations. In Table 3 (A) the random selection is repeated 1000 times and the overall classification rates are averaged. We see here that the Support Vector Machine classifer provides a substantial gain over the other methods we have considered.

We also investigated the impact of automatic transcription (output from the speech recognizer) on the performance of the classifiers. Because of the computational resources needed to automatically transcribe all 743 conversations, we provide results in Table 3B and 3C for a single subset of 75 test conversations (the classifiers are trained using the 668 remaining transcribed conversations). Classification error rates on manually transcribed conversations for this single test set are shown in Table 3 (B). We provide these results for comparison purposes with the more exhaustive simulations shown in Table 3 (A). Next, we automatically transcribed the 75 test conversations using our baseline Switchboard recognition system to obtain the resulting classification error rates shown in Table 3 (C). We see that the classifier performance is not significantly degraded by the use of automatic speech recognition transcriptions. For example, the error rate for the SVM classifier increases slightly from 24.1% to 27.8% error.

Table 3. *Call Type Classification Error Rates (%).*

	(A)	(B)	(C)
Language Model	33.9	27.8	46.8
Naive Bayes	31.8	30.4	32.9
TFIDF	28.0	26.6	31.6
Support Vector Machine	24.0	24.1	27.8

Transcribing data to build call type classifiers is extremely expensive since human labeling and transcription of the calls are required. One question to be answered is how the call-type classification error rate is impacted based on the number of transcribed conversations. Second, we wish to understand whether automatically transcribed data can be used to boost classification performance.

An illustration of these scenarios is shown in Figure 2. We plot call-type classification error rate along the y-axis vs. the number of manually transcribed conversations along the x-axis. In line (A) we see that the classification error rapidly drops until approximately 300 conversations are hand-transcribed.

In order to boost classification performance while using less manually transcribed data, we generated automatic transcriptions for the untranscribed data portion of the call center corpus. For each automatically transcribed conversation, we manually label the ASR output text file into one of 6 categories (call-types). We consider this labeling effort as “lightly” supervised training since it is less expensive and time consuming to classify calls using text output from a speech recognizer (6 minutes classification time per hour of speech) compared to classifying the same calls using manual transcripts (3 minutes classification time + 3 hours transcription time per hour of speech). We combine this lightly supervised training data with increasing amounts of manually transcribed data and plot the results in line (B). Here, we have incorporated only 200 automatically transcribed (but hand-classified) conversations into the training pool for our call-type classifier. The 200 conversations were selected from a possible pool of 329 untranscribed conversations based on acoustic confidence averaged over the duration of the call. We see that with just a very limited amount of untranscribed data one can improve the system classification performance substantially. These results suggest that a combination of manual transcription in tandem with lightly supervised call labeling can allow for rapid system portability.

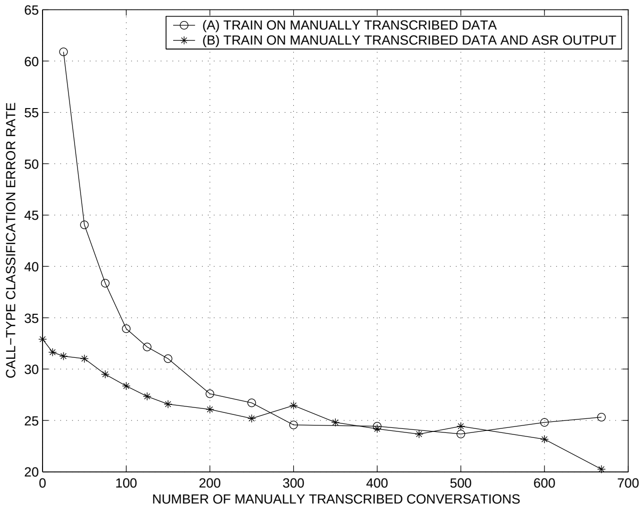

Fig. 2. *Call-type classification vs. number of manually transcribed conversations.*

6.2. Call Transcription

Our test set for recognizer evaluation consists of 24 agent sessions and 24 caller sessions comprising 1.5 hours of audio data (30 minutes voice activity). Word error rates for call transcription are shown in Table 4. We see that the baseline system using Switchboard acoustic and language models (Table 4A) has a 53.3% WER for the caller data and 54.2% for the agent data. Retraining the acoustic model by incorporating untranscribed data from the corpus lowers the WER by 1.1% absolute for the caller channel and by 4.8% absolute for the agent channel. The larger gain for the agent call-sides is most likely due to the fact there are a relatively few number of agents (˜20). Interestingly, we see a small reduction in error on the agent data after training the language model from text automatically derived from the untranscribed portion of the corpus. Note that for this case, Table 4C, the new vocabulary size (8k words) is much smaller than that of the original Switchboard language model (34k words). We see further gains by including task-specific webpage material from the IT Service Center website (Table 4D). Including word-posterior confidence into acoustic retraining provides marginal impact on word error rates (Table 4E).

Finally, we compare the results from unsupervised training with the same baseline system which uses the Switchboard acoustic models and a class-trigram language model derived from the transcribed text portion of the call center corpus (Table 4F). This language model has a 6k vocabulary size with 10 manually determined word-classes. The final word error rate from complete system retraining (Table 4G) is 43.5% for the caller conversations and 27% for the agent conversations. We point out that there still exists a large difference between unsupervised training and the same system retrained from human transcribed material. The major difference appears in modeling and characterization of the language model for the call center environment.

Table 4. *Word Error Rate (%) for Call Center test set. Legend: Switchboard (SWB), Untranscribed Data (UNT), Confidence (Conf), Transcribed Data (TR).*

	Acoustic Model	Language Model	Lex	Caller	Agent
(A)	SWB	SWB	34k	53.3	54.2
(B)	SWB+UNT	SWB	34k	52.2	49.4
(C)	SWB+UNT	UNT	8k	52.9	48.6
(D)	SWB+UNT	UNT+WebPage	7.5k	50.7	44.1
(E)	(D)+Conf	(D)+Conf	7.5k	50.1	43.6
(F)	SWB	TR	6k	46.1	38.3
(G)	SWB+TR	TR	6k	43.5	27.0

7. CONCLUSIONS

In this paper we described recent experiments in call-type classification and acoustic modeling for speech recognition in the context of a new telephone-based call center corpus. We decribed the collection of a new audio corpus from a IT Help Desk located at the University of Colorado. We described our analysis of the corpus in terms of call types and compared four methods for automatic classification of calls into 6 broad classes. It was shown that the Support Vector Machine classifier provides the best classification performance (error rate of 24% using transcribed conversations and 27.8% using automatic transcriptions from a speech recognizer). We provided an initial investigation of methods for unsupervised training of the call center acoustic and language model. Results show improvements over a baseline system trained from the Switchboard corpus, but there still exists a larger margin for improvement given the word error rates of the same system trained on transcribed conversations.

8. ACKNOWLEDGEMENTS

This research was partially sponsored by Avaya Inc. via research award W0711119CWS. The authors would like to thank Al Literati, Robert Steiner, Wu Chou, and Valentine Matula of Avaya for contributions related to hardware and data collection set up and project support. The authors would also like to thank Taylor Struemph of CSLR for aiding in the transcription of the call center corpus.

9. REFERENCES

[1] A.L. Gorin, G. Riccardi and J.H. Wright, “How May I Help You?”, *Speech Communication*, vol. 23, pp. 113–127, 1997

[2] R. Iyer, H. Gish, D. McCarthy, “Unsupervised Training for Natural Language Call Routing”, *ICASSP*, Orlando, Florida, vol. IV, pp. 3900–3903, 2002

[3] P. Natarajan, R. Prasad, B. Suhm and D McCarthy, “Speech-Enabled Natural Language Call Routing: BBN Call Director”, *ICSLP*, Denver, Colorado, pp. 1161–1164, 2002

[4] Lori Lamel, Jean-Luc Gauvain, Gilles Adda, “Unsupervised Acoustic Model Training”, *ICASSP*, Orlando, Florida, vol. I, pp. 877–880, 2002

[5] Lori Lamel, Jean-Luc Gauvain, Gilles Adda, “Lightly supervised and unsupervised acoustic model training”, *Computer Speech and Language*, vol. 16(1), pp. 115–129, 2002

[6] Dilek Hakkani-Tűr, Giuseppe Riccardi and Allen Gorin, “Active Learning for Automatic Speech Recognition”, *ICASSP*, Orlando, Florida, vol. IV, pp. 3904–3907, 2002

[7] Jason D. M. Rennie and Ryan Rifkin, “Improving Multiclass Text Classification with the Support Vector Machine”, Massachusetts Institute of Technolgy, *AI Memo AIM-2001-026*, 2001

[8] McCallum, Andrew Kachites, “Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering”, *http://www.cs.cmu.edu/ mccallum/bow*, 1996

[9] Kadri Hacioglu, Wayne Ward, “A Concept Graph based Confidence Measure” , *ICASSP*, Orlando Florida, vol. I, pp. 225–228, May 2002

[10] Bryan Pellom, Kadri Hacioglu, “Recent Improvements in the CU Sonic ASR System for Noisy Speech: The SPINE Task”, *ICASSP*, Hong Kong, vol. I, pp. 4-7, April, 2003

[11] Timothy Kuhn, Michele Jackson, “Accomplishing Knowledge: A Communicative Model of Knowledge Applied to a Call Center”, *submitted to Communication Theory*, March 2003