Using Semantic Similarity In Automated Call Quality Evaluator For Call Centers

Ria A. Sagum, MCS

Abstract: Conversation between the agent and client are being evaluated manually by a quality assurance officer (QA). This job is only one of the responsibilities being done by a QA and particularly eat ups a lot of time for them which lead to late evaluation results that may cause untimely response of the company to concerns raised by their clients. This research developed an application software that automates and evaluates the quality assurance in business process outsourcing companies or customer service management implementing sentence similarity. The developed system includes two modules: speaker diarization, which includes transcription and question and answer extraction, and similarity checker, which checks the similarity between the extracted answer and the answer of the call center agent to a question. The system was evaluated for Correctness of the extracted answers, and accurateness of the evaluation for a particular call. Audio conversations were tested for the accuracy of the transcription module which has an accuracy of 27.96%. The Precision, Recall and F-measure of the extracted answer was tested as 78.03%, 96.26% and 86.19% respectively. The Accuracy of the system in evaluating a call is 70%.

Index Terms: Automatic Transcription, Speaker Diarization, Text Similarity checking

1. INTRODUCTION

Quality assurance (QA) is the planned and systematic activities implemented in a quality system so that quality requirements for a product or service will be fulfilled. In a business process outsourcing company or customer service management, quality assurance specialist participates in customer and client listening programs to identify customer needs and expectations. This is a critical part of a company to maintain its current service standard. At the present time companies have customer service department to provide fast and precise solutions for the clients. Since QA are responsible in different activities there is a problem of untimely evaluation result for every call transactions leading to lateness of the response of the company for a particular concern. The researcher opted to develop an application software that will automate the quality assurance job that will evaluate the transaction for every call.

2. RELATED WORKS

Quality assurance (QA) should be the cornerstone of the call center management efforts. This is because optimizing QA practices will help to enhance the quality of the service your team provides to their customers, increase their efficiency and reduce wasteful spending [1]. According to a blog from Genesys, entitled Quality Management: Improve Call Center Quality Assurance and Agent Performance by Completely Understanding Every Conversation, quality management allows you to automatically analyze every conversation, measure agent skills against objective criteria and gain a true understanding of every agent's performance [2]. Call center quality assurance programs ensure that your customers receive a consistent standard of service when they contact a call center or when a call center agent contacts them [3].

Automation of quality assurance program is not an easy task. One technology that is showing particular promise is a computer's ability to recognize human speech or Speech-to-Text (STT) [4]. Current speech analytics technology boasts accuracy significantly greater than 80 to 90 percent. With improved accuracy, speech analytics have been working diligently to improve the speed at which results are delivered [5]. Most of the existing studies that is testing for a quality is usually quality testing for a product. There were only a few studies that presented methods and techniques in order to test the quality level of a call. On Stepanov paper [6], it automatics summarized spoken conversation in terms of factual descriptors and abstractive synopses that are useful for quality assurance supervision in call centers. While on Pallotta, V. et.al study [7], they have presented a new approach to Call Center Analytics based on Interaction Mining, contrasting Text Mining, which is currently used in Speech Analytics in order to provide useful insights for enhancing Call Center Analytics to a level that will enable new metrics and key performance indicators (KPIs) beyond the standard approach. The paper entitled Automated Quality Monitoring for Call Centers Using Speech and NLP Technologies [8] had presented an automated system for quality monitoring in the call center. They proposed a combination of speech recognition, maximum entropy classification based on ASR-derived features, and question answering based on simple pattern-matching. The system can either be used to replace human monitors, or to make them more efficient.

3. METHODOLOGY

The diagram below (Fig. 1), shows the design of the developed evaluator system. The input contains the audio conversation. The file that contains the conversation will be synthesized. The synthesized conversation converted as a text file will become the input of the next module called Diarization module.

- *Ria A. Sagum is currently an Associate Professor in the Department of Computer Science at the College of Computer and Information Sciences at the Polytechnic University of the Philippines, Philippines,*
- *E-mail: rasagum@pup.edu.ph*

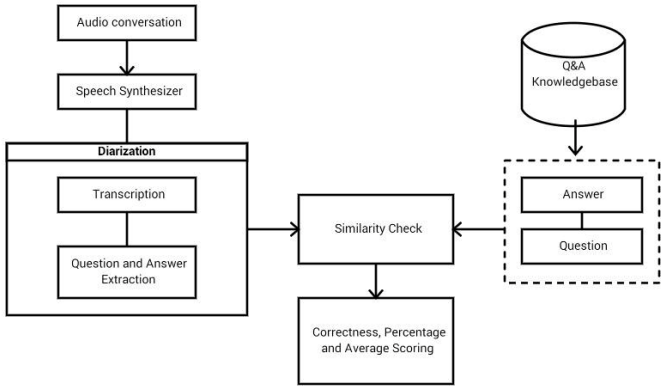

Figure 1: *System Architecture of the Call Evaluator System*

Diarization, contains two processes the Transcription and Question and Answer Extraction. The processes are defined as follows:

A. Speaker Diarization and Transcription

Using LIUM SpkDiarization tool, a software dedicated to speaker diarization [9], the input audio signal is analyzed for speaker segmentation and clustering. Using FFMPEG, a tool for handling multimedia data [10], every speech clusters will be extracted to a temporary wav file. Every wav file will then be transcribed using Sphinx4, a state-of-art HMM-based speech recognition system being developed on open source. [11]

B. Question and Answer Extraction

The Question and Answer Extraction module requires the transcriptions of call that was converted from voice to text**.** This module accepts tagged statements (Question, Complaint, or Declarative). The tagging of the statements are being done by the system.

Question Tagging

To tag statement as question, POS tagger were used to identify if the sentence has a question class - Who, What, Where, When, Why, How, etc. If the sentence doesn't have question class, it will automatically tag as Declarative.

Complaint Tagging

To tag a statement as complaint, SentiWordNet [13] was used. Every word that has sense value - Noun, Verb, Adverb, Adjective, in the sentence will be evaluated and get the senti value. After getting each of the values, it will be totaled. If the total value is less than zero, the statement will be considered as complaint, else, Declarative.

Only those statements that are tagged as Question or Complaint will undergo **Answer Extraction**. Using WordNet, synonyms of words that has sense in the statement - noun, adjective, verb and adverb [12] will be considered to determine if the candidate answer (Operator's Statements) is the right answer to that particular Question/Complaint. If the system failed to find synonyms from the candidate answers, immediate response statement from the operator will be considered as the answer to that particular Question/Complaint. As for the Declarative statements, immediate answer of the Operator will be considered as the answer to the caller's statement.

Similarity Check

This module will first identify the most similar question from the input to a file consisting of the most common questions being answered and its corresponding answers. It was revealed that call center agents' replies for every customer's inquiries are based to the set of common questions and answers given by the company. The system will use this file to generate the correct answer for each questions. Text Mining was done to determine all possible answer related to the question. The system after generating the answer will evaluate the call/transaction by checking the similarity of the agent's answer and the one generated by the system. The semantic similarity method was used to determine the sentence similarity between the statements. It made use of the set of words found in the statements being compared. The semantic calculation was done by using a Semantic Similarity tool [14]. The categories came from a new dictionary which is based from the English Open Word List (EOWL) of words, while its semantic similarity for each word is calculated from the DISCO's semantic similarity. The output of the evaluation is the overall similarity for the entire conversation which then can be used to assess the success of the call/transaction.

4. RESULTS

To evaluate the performance of the developed system Precision, Recall and F-measure was used by the researcher. To test the accuracy of the Speaker Diarization module which consists of Transcription and Diarization of the system, the researchers used a command in Linux to train the wav files, then it will output the words it generated and will later compare it to the training data. Lastly, the system will compute for its accuracy. The system computed 27.96% of accuracy out of 208 audio conversations tested for Speaker Diarization module as shown below. With this result the researcher opted to use a tool that will convert a text to speech to continue working on the main point of the research which is the capability of an application system to automatically evaluate the success of a call. In evaluating the accuracy of the Question and Answer Extraction module, the researchers used F-measure. This measures considers both the precision (P) and the recall(R) of the test to compute the score. The F-measure score can be interpreted as a weighted average of the precision and recall. Based from the tagged questions, the answers were extracted from the transcription as tagged answers. This will serve as the call center's answer to a question. The accuracy of tagging the answer for every questions was then computed by identifying the TP, TN, FP, FN based the list of questions and answers. The module was tested using 17 audio conversations with an average of 10 statements for every conversations. For the computation of F-measure, the following formulas were used:

F-measure=2 (Precision*Recall)/(Precision+Recall)

Equation 1 Formula for F-measure

Precision= tp/((tp+fp))

Equation2 Formula for Precision

Recall= tp/((tp+fn))

Equation 3 Formula for Recall

Where:

- True Positive: correctly tagged as answer/question
- False Positive: tagged as answer but not

- False Negative: Results missing information
- True Negative: No answers were given

Table 1: *Summary of Findings for Precision, Recall, F-measure of the system for the Question and Answer Extraction*

	Precision	Recall	F-measure
Answer /Question Tagging	78.03%	96.26%	86.19%

Table 1shows the computed results that leads to a 78.03% for the Precision rate of the Question and Answer Extraction module, while 96.26% for the Recall Rate, 86.19% for the F-measure and has an Accuracy of 75.74%. One hundred three (103) were identified as True Positive (TP) from the tagged questions and generated answers of 17 audio conversations, Twenty nine (29) for the False Positive (FP), Four (4) for the False Negative and none of the results identified as True Negative (TN). It can be easily seen that the Recall Rate has the highest result. For the Similarity Check module, the percentage for the Question and Answer was computed. Based on the output of Question and Answer Extraction module, it will be the input to be compared to the similarity checker, then the system itself outputs the computed percentage of similarity per category. The computed percentage was interpreted using Overall Similarity Index [15]. **(See table 2)** As shown in the table 2, the Percentage of Similarity was tested using the audio conversations. From the computed percentage of similarity per category, the results were average by dividing the total number per category to One hundred thirty six (136) total test data of 17 audio conversations. The Similarity Checker Module has a 76.90% of similarity interpreted as Very High Similarity for the Question, while there is 53.55% similarity for the Answer which is interpreted as High Similarity.

Table 2: *Verbal Interpretation of Similarity Accuracy*

Rating	Level of Similarity
0-24%	Similarity Not Occurred
25%-49%	Average Similarity
50%-74%	High Similarity
75%-100%	Very High Similarity

Based on the similarity percentage accumulated in Similarity Checker module, the values are evaluated if it is considered similar or not similar. To compute for the Threshold, the researchers get the mean of all the True Positive results and accumulated 51.75%. (See table 3) It will be the basis to interpret if the question and answer is similar to what the BPO Companies QA is. It is considered similar if the similarity percentage is greater or equal to the threshold, otherwise, not similar.

Table 3: *Threshold*

Threshold	51.75%

The accuracy of the developed system based on the capability to evaluate every call was computed as:

Accuracy = The. No. of correct evaluation/The no. of Total Evaluation

It was gathered that from the tagged files that there was a total of 272 question and answers statements and the accuracy of the evaluation of the system in the evaluation is 70% (See Table 4).

Table 4: *Accuracy of the Call Evaluator*

Correct Eval.	Incorrect Eval	Accuracy
191	81	70%

5. CONCLUSION AND RECOMMENDATION

This project aims to develop an application software that automates the evaluation of a call by quality assurance officer in business process outsourcing companies or customer service management. The findings and evaluation result of the project proved that the system can automate and evaluate Quality Assurance in BPO Companies. In conclusion, the analysis shows that there is a 27.96% accuracy in Transcription module, while there is 78.03% for Precision Rate, 96.26% for Recall Rate, and 86.19% for F-measure for the Accuracy of the Question and Answer Extraction module. The computed Accuracy for call evaluation was 70%, which looks promising considering that this is a new field. Other researchers may expand the system further by:

a. Improving the Diarization module
b. Having different set of training data for the LIUM Diarizer and Sphinx.
c. Using better Similarity Checker tool aside from Semantics tool.
d. Implementing Pattern Recognition for the Answer Extraction.

6. ACKNOWLEDGEMENT

The researcher would like to thank the following persons who extended a helping hand for the success of this study. The QA group of PUP-SIG-NLPconsists of the following members: Casielyn Angelia D. Alivio, Jerome Basco, Elgie Candelario, Francis Gann Claveria, Rody Christian Dulfo, Mark Lawrence Estalilla, Mark Rowell Gayon, Jasmine Golperic, Gerald Gumabon, Rowsette Dorothy D. Llanes, Marionne Nicole G. Mamanta, Jamie Paul Medrano, Ma. Ana Casandra Miraballes, James Thomas S. Regpala, Jerome Santiago for their help while implementing the project.

7. REFERENCES

[1] Talkdesk. Quality Assurance Best Practices in the Call Center. Internet: https://www.talkdesk.com/blog/call-center-management/quality-assurance-best-practices-in-the-call-center/, 2013 [March 2016].

[2] Genesys. Quality Management: Improve Call Center Quality Assurance and Agent Performance by Completely Understanding Every Conversation. Internet: http://www.genesys.com/platform-services/workforce-optimization/quality-management, 2015 [March 2016].

[3] Linton, I. Demand Media, About Call Center Quality Assurance Programs, Hearst Newspaper, 2015. Internet: http://smallbusiness.chron.com/call-center-quality-assurance-programs-39627.html, 2015 [March 2016].

[4] Hwabgbo, H. Mining Customer Insights with Speech-to-Text Technology. Internet: http://usblogs.pwc.com/emerging-technology/mining-customer-insights-with-speech-to-text-technology/, 2014 [March 2016].

[5] Vreede, S. V. What is Speech Analytics? Call Center and Business Applications. Internet: http://www.smartcustomerservice.com/Articles/What-Is/What-is-Speech-Analytics-Call-Center-and-Business-Applications-98113.aspx, 2014 [March 2016].

[6] Stepanov, E. et.al., “Automatic Summarization of Call-Center Conversations”

[7] Pallota, V. et.al., “Interaction Mining: The new frontier of Call Center Analytics”

[8] Zweig, G., “Automated Quality Monitoring for Call Centers Using Speech and NLP Technologies”, Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 292–295, June 2006.

[9] "FFmpeg License and Legal Considerations". ffmpeg.org. Retrieved 2012-01-04.

[10] P. Lamere, P. Kwok, W. Walker, E. Gouvea, R. Singh, B. Raj, P. Wolf, "Design of the CMU Sphinx-4 Decoder," cmusphinx.sourceforge.net, 2003 (USA).

[11] S. Meignier, T. Merlin, “LIUM SpkDiarization: An Open Source Toolkit For Diarization,” in Proc. CMU SPUD Workshop, March 2010, Dallas (Texas, USA).

[12] George A. Miller, “WordNet: A Lexical Database for English”, Communications of the ACM Vol. 38, No. 11: 39-41, 1995.

[13] A. Das and S. Bandyopadhyay, “SentiWordNet: Dr Sentiment Knows Everything!”,ACL/HLT Demo Session, Pages 50-55, June2011.

[14] Sourceforge, “Calculate Semantic Similarity”, Internet: https://sourceforge.net/projects/semantics/, August 23, 2012 [March 2016].

[15] E-Learning Tees. “The Overall Similarity Index”, Internet: https://eat.scm.tees.ac.uk/bb8content/resources/recipes/interpretTurnitin.pdf, August 2007 [March 2016].