March 08, 2019

# MACHINE LEARNING-BASED GENERATION OF CONTEXTUAL FREQUENTLY ASKED QUESTIONS FOR PROBLEM ANALYSIS IN SUPPORT ENGAGEMENTS

Chidambaram Arunachalam

Hazim Dahir

Alka Luqman

# MACHINE LEARNING-BASED GENERATION OF CONTEXTUAL FREQUENTLY ASKED QUESTIONS FOR PROBLEM ANALYSIS IN SUPPORT ENGAGEMENTS

AUTHORS:
Chidambaram Arunachalam
Hazim Dahir
Alka Luqman

## ABSTRACT

Techniques are described herein to automatically generate Frequently Asked Questions (FAQs) during the initial problem analysis of a given issue. The conversation transcripts of successfully resolved recent cases are analyzed using a supervised machine learning model to predict the initial problem analysis phase of a conversation and to extract FAQs asked during this phase. The extracted FAQs are displayed to the customer at the time the case is opened, thereby accelerating the problem solving process and augmenting the support engineer's troubleshooting capabilities.

## DETAILED DESCRIPTION

The key to quickly solving a problem is first clearly understanding the problem. Oftentimes, support engineers in networking, manufacturing, and many other industries spend several hours troubleshooting an issue without asking the right set of questions to better understand the problem. In one example, an engineer troubleshoots a high-availability issue in which the standby router periodically loses synchronization with the active router and reloads intermittently. The engineer may ask whether there is any packet loss or interface error. If the customer indicates that no issues have been observed in the local interfaces and everything else appears to be functioning normally, the engineer may enable debugs to troubleshoot the problem and wait for the next occurrence of the problem. The missing questions that could have helped solved the problem include "What is the end-to-end network path between the two routers?"; "Are there any packet drops in any of the network interfaces across the end-to-end network path?"; etc.

Today, questionnaires are systematically built by humans who have experience troubleshooting a given problem type. Human experts in the given technology field ask the right set of questions and solve problems faster. However, the experts have to manually document these questions in order to share their knowledge with their peers, and the peers have to find the right article to identify the right questionnaire.

Described herein is a mechanism that automatically generates contextually relevant Frequently Asked Questions (FAQs) for a given problem type by analyzing conversations that correspond to tickets in which the given problem type was successfully resolved in the recent past. The automatically generated questionnaire may then be presented to the customer and answered by the customer at the start of the support engagement process such that the support engineer has a very clear understanding of the problem and may start the troubleshooting process in the right direction from the beginning. The automatically generated questionnaire may also be presented to new support engineers joining the workforce to augment their troubleshooting capabilities.

Every support engagement (e.g., ticket, case, etc.) has a wealth of data including the customer's description of the symptoms as well as data captured prior to and during the issue resolution process. In addition to the diagnostics data, another highly valuable data set is the interaction between the customer and the support engineer (case owner). The conversation between the customer and engineer goes through a series of phases during the lifecycle of a case (e.g., initial problem analysis, troubleshooting, resolution, monitoring).

Typically, support engineers review the data provided by the customer and then provide an initial set of questions to better understand the problem. This is often referred to as initial problem or situation analysis. For lower-severity issues, the questions and corresponding answers exchanged across email help the engineer identify next steps for troubleshooting. Oftentimes, this initial exchange may require several minutes or even hours, resulting in delayed time to understand, troubleshoot, and resolve the problem. Also, if the support engineer is new to the problem space, he/she must manually reference past cases and/or articles to understand the right set of questions to be asked in order to correctly solve the reported issue.

Customers, partners, and support engineers have now started to use persistent virtual collaboration spaces as the primary communication channel for a given support engagement. The virtual space is associated with the support case and accelerates the problem solving process by enabling support engineers to ask initial questions and obtain answers in real-time while continuing to ask questions to the void until the reported problem is clearly understood. The transcript of the virtual space conversation is easily retrievable from the communications platform using Application Programming Interfaces (APIs). The knowledge gained from the conversation analysis of transcripts of past cases that were successfully resolved provides techniques to automate generation of initial questions. This automation approach minimizes the possibility of taking a wrong troubleshooting path (by a new support engineer) and avoids any delays in understanding the problem by presenting the generated questions to the customer immediately after the case is opened and requesting them to answer while waiting for a support engineer to be assigned to the ticket.

Figure 1 below illustrates three example steps.

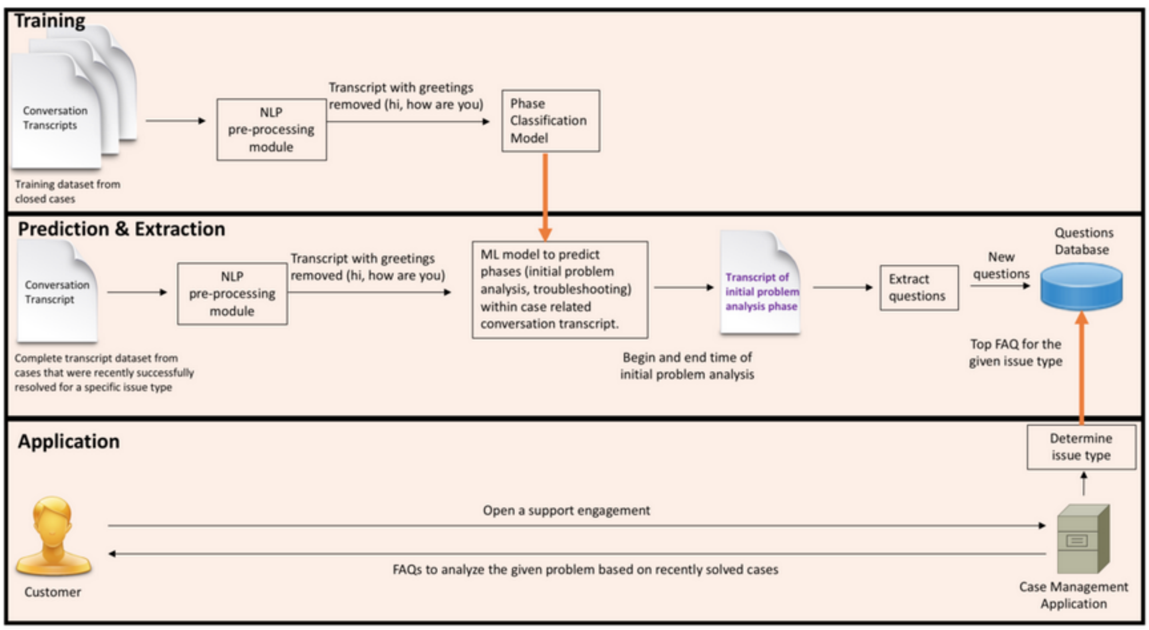


Figure 1: Automatic Generation of Frequently Asked Questions during initial problem analysis

The first step is to train a classification Machine Learning (ML) model to predict whether a given message belongs to initial problem analysis phase. The second step ~~is~~ involves prediction and extraction. The classification label may be predicted for each message within the conversation transcript of a recent successfully resolved case. This is used to detect the timestamp at which the conversation transitions from the "initial problem analysis" phase to the "troubleshooting" phase in the issue resolution process. This module also extracts questions from the messages that are part of the initial problem analysis phase. New questions are added to the questions database. If the question already exists, the frequency counter for that question is incremented.

In the third step, when a customer opens a case, the case management application identifies the issue type, looks up the top FAQ(s) that were auto-generated in the previous step, and displays them to the customer user. The customer answers these questions based on the information available to accelerate the problem solving process.

The training data source includes the conversation transcripts which are extracted from the collaboration virtual space used by the customer and the support engineer to communicate about the issue. This set of transcripts from closed cases forms the data corpus on which a neural network model is trained to predict when the phase changes from Initial Problem Analysis (P1) to Troubleshooting (P2), etc. in a case's lifecycle. A supervised classification approach may help solve this.

The data preparation phase involves cleaning the extracted data and imputing missing information. This entails removing bot-generated messages, removing columns which are not significant contributors to model-building, and detecting and removing greeting related messages such as "Hello," "Good Morning," etc. The greeting removal is implemented using a Natural Language Processing (NLP) intent classification model, trained on language models available in the open source domain. The sentences in the transcripts are then tokenized and converted to vector notation using word embedding techniques. The scoring mode used is Term Frequency - Inverse Document Frequency (TF-IDF) for each word in the sentence. Thus, each human-uttered sentence from the dataset is now encoded as a 15,000-element vector, where each element indicates the presence/absence of that word. The absence of a word causes the corresponding element to be assigned a value of zero, and the presence of a word is indicated by its corresponding TF-IDF score in that element. This completes the data transformation stage.

The model itself is a neural network of multiple dense layers with 512 neurons in each layer. A dropout metric may be used at each layer in order to avoid overfitting. A cross entropy loss function is used to reduce misclassifications between the predicted phase and the actual labeled phase.

The successfully resolved cases that correspond to a given issue type in the recent past are first identified. The same preprocessing steps as in the training phase are applied to the conversation transcript of the identified cases. Now, the set of sentences in the transcript is converted into a matrix. Each row in the matrix is the vector notation of the sentence over the 15,000-element vocabulary. This matrix forms the input to the neural network classification model developed in the "Training" step.

The model predicts each message within the given transcript as belonging to "initial problem analysis," "troubleshooting phase," etc. The system identifies the change in phase transition within the transcript and outputs the timestamps of the beginning and ending messages of the initial problem analysis phase. Figure 2 below illustrates an example in which a conversation transcript has 100 messages. The first ten messages are classified as messages that belong to the initial problem analysis phase and the rest are classified as the troubleshooting phase. In this case, the timestamps of messages 1 and 10 are provided as outputs of this module.

| Message Timestamp | Message Content | Classification Label |
|---|---|---|
| T1 | M1 | P1 |
| T2 | M2 | P1 |
| T3 | M3 | P1 |
| T4 | M4 | P1 |
| T5 | M5 | P1 |
| T6 | M6 | P1 |
| T7 | M7 | P1 |
| T8 | M8 | P1 |
| T8 | M9 | P1 |
| T10 | M10 | P1 |
| T11 | M11 | P2 |
| ...... | ... | ... |
| T100 | M100 | P2 |

Figure 2: Phase prediction for each message within the transcript

The beginning and ending timestamps are then used to generate a sub-transcript which contains only the messages that belong to the "initial problem analysis" phase (P1). Each of these messages are then processed to determine whether they are a question.

If it is a question, it is added to the "Questions Database". The Questions Database is a collection of user questions organized into clusters as shown in Figure 3 below, each cluster representing conceptually similar questions, which may be the many different ways users ask the same question with regard to a specific type of issue. All questions will be converted to vector notation using word embedding techniques. This will ensure that similar words (and thus questions) will occupy close spatial positions. Similarity between questions may then be calculated using distance metrics like cosine similarity.

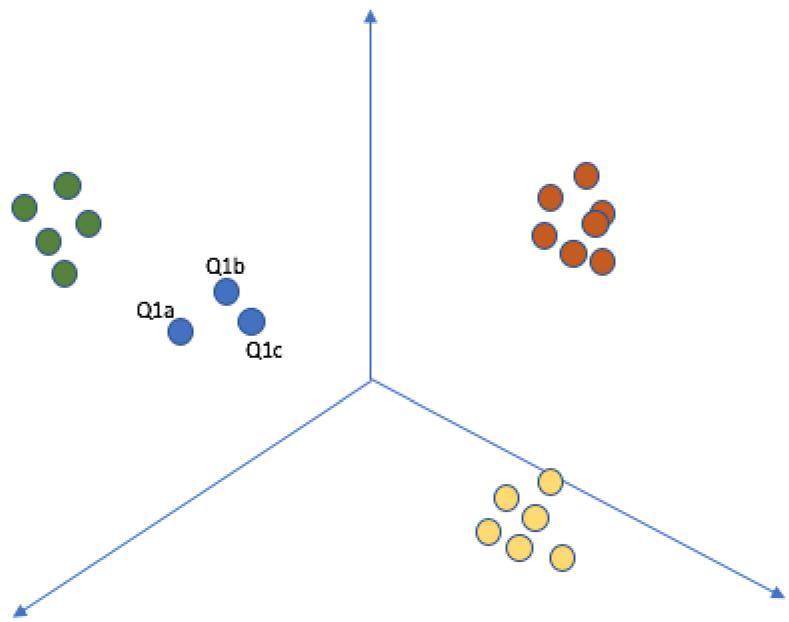


Figure 3: Question Collection Visualization

A collection may be a set of conceptually similar questions. For each collection, the different ways a user can ask the same question (size of the cluster) may be tracked along with the number of times each question has been asked (using a frequency counter), as illustrated in Figure 4 below.

| Similar Question Collection | Collection Size | Frequency Counter |
|---|---|---|
| {Q1a, Q1b, Q1c } | 3 | 100 |
| {Q2a, Q2b, Q2c , …} | 5 | 50 |
| {Q3a, Q3b, Q3c, … } | 6 | 20 |
| {Q4a, Q4b, Q4c } | 3 | 90 |
| {Q5a, Q5b, Q5c, … } | 10 | 95 |
| {Q6a, Q6b, Q6c, … } | 4 | 5 |
| {Q7a, Q7b, Q7c, … } | 6 | 40 |
| {Q8a, Q8b} | 2 | 30 |
| {Q9a, Q9b, Q9c, … } | 8 | 1 |
| {Q10a, Q10b, Q10c, … } | 5 | 20 |

Figure 4: Questions database

For example, with reference to the first collection from the Question Cluster Database shown in Figure 4, the users may be asked questions regarding the version of the phone in three different ways (e.g., Q1a: "What is the version of your Internet Protocol (IP) phone?"; Q1b: "What software version are you running on the phone?"; and Q1c: "Which Operating System (OS) version is installed on the phone?"). Each question when converted to its corresponding word vector will occupy a point in vector space, as represented by the labeled points in Figure 3. The cluster size (three) represents the variety in the number of ways the same question is asked. There may be cases in which the question was framed in the same manner by different people. This may be recorded in the frequency counter for each question. The frequency counter shown in Figure 4 is a cumulative metric. In this example, the frequency counter is 100, which is actually maintained on a per question basis as a set (e.g., {35,50,15} for {Q1a,Q1b,Q1c}, indicating that question Q1a was asked twenty-five times, Q1b fifty times, and Q1c fifteen times.

The customer uses a case management application to submit a new case. The case management application processes the problem description, determines the issue type (the case submission form may also have a field for issue type) and then identifies the top FAQs during problem analysis of this specific issue type. This "top frequency" metric is a combination of the cluster size and frequency counter as described in the "Extraction" step above. These FAQs are then presented to the user as shown in Figure 5 below. The customer user answers the questions based on the current situation. This information is thus readily available to the assigned engineer to quickly understand the problem and begin with further questions or proceed with the troubleshooting process. The support engagement may also be triggered from a network assurance system. Similar customer experience may be provided through the network assurance system website.

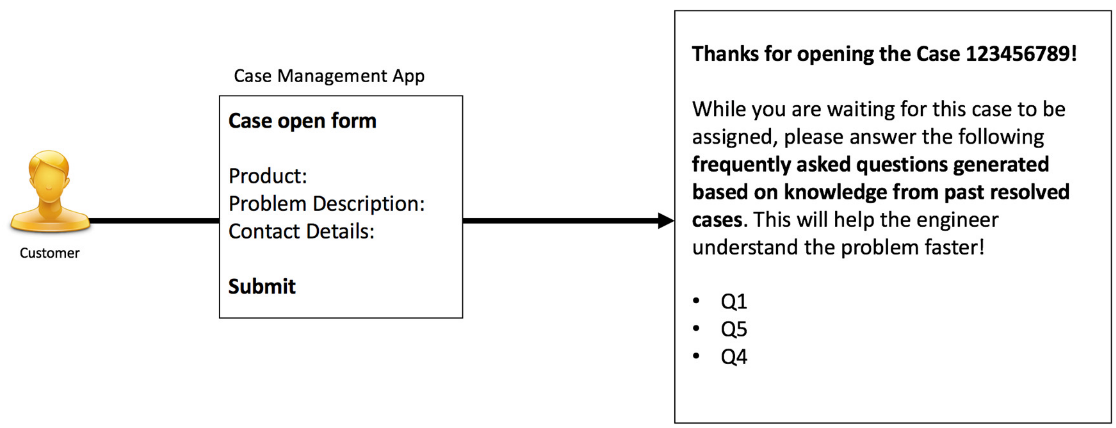


Figure 5

In summary, techniques are described herein to automatically generate FAQs during the initial problem analysis of a given issue. The conversation transcripts of successfully resolved recent cases are analyzed using a supervised machine learning model to predict the initial problem analysis phase of a conversation and to extract FAQs asked during this phase. The extracted FAQs are displayed to the customer at the time the case is opened, thereby accelerating the problem solving process and augmenting the support engineer's troubleshooting capabilities.