Interspeech 2022
18-22 September 2022, Incheon, Korea

Evaluation of call centre conversations based on a high-level symbolic representation

Leticia Arco[1], *Carlos Mosquera*[1], *Fabjola Braho*[1], *Yisel Clavel*[1,2], *Johan Loeckx*[1]

[1]Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium
[2]Universidad de Holguín, Holguín, Cuba
larcogar@vub.be, yclavelq@uho.edu.cu, jloeckx@vub.be

Abstract

We present a demo that illustrates the performance of our system to analyse and evaluate call centre conversations. Our solution can be used at different stages of the quality feedback loop. The high-level symbolic representation developed on the context-based intent recognition core module allows for detecting fine-grained reasons for quality assurance problems and going in-depth qualitative analysis of how agents and customers interact. We illustrate the evaluation and insights of real-life conversations provided by a Belgian call centre. Participants can interact with the demo by playing with call annotation, recommendations, and diverse parameters.

Index Terms: call monitoring, intent recognition, human-human interaction, recommender systems

1. Introduction

Quality Monitoring (QM) of conversations is a critical task in call centres. Its main issues are: monitoring at least doubles the cost of each call, only a few calls are listened to and therefore exploited, random sampling can cause listening to uninteresting calls, and the considerable bias among the QM supervisors can lead to high disagreement on the assessment.

The need for automatic tools allowing to monitor and mine calls is high. It contributes to improving customer insight and therefore develops loyalty. Solutions exist that can evaluate call centre conversations [1]–[3]. However, their main limitations are: offer a superficial analysis of indicators, analyse few basic KPIs, only support manual evaluation, failing to explain the provided evaluation, or just extracting the key utterances. Most metrics are not aimed at measuring whether or not the solution to the posed problem was effective, nor if the Quality Assurance (QA) checklist was fulfilled. Semantic analysis of conversations is not performed (e.g., an agent could have good scores in basic indicators and not solve the customers' problems). Besides, the evaluation methods do not fully take advantage of the knowledge extracted from other call centre tasks and vice versa.

In this demo, we showcase a system that completes a call QA checklist based on a high-level symbolic representation of conversations, not only superficial statistics/metadata. This representation, based on the intent recognition core module, allows gaining insight into why call quality is low, detecting fine-grained reasons for the QA problems, offering a more robust solution, and providing results that serve as a basis to support other tasks besides QA. The aim of our demo is threefold: (1) to describe the new model for evaluating dialogues based on their semantic representations, (2) to show the context influence on intent recognition, and (3) to highlight how the extracted patterns impact recommendations.

2. System components

Our proposal completes a call QA checklist by following five stages: dialogue reconstruction, speaker identification, intent detection, symbolic representation, and call evaluation.

Our demo offers the interaction with the developed DialogueKit web application, which allows us to identify speakers and annotate dialogues, among other functionalities. The speaker identification solution is distinguished by the post-classification step, where the best-classified speaker influences how to classify the other one. The flexible dialogue annotation is an added value of our solution because intents do not have to be strictly associated with an utterance, i.e., intent examples can go through more than one utterance, coincide with or be part of an utterance, as shown in Figure 1.

Figure 1: *Excerpt of the DialogueKit interface.*

Recognizing intents is the core of the obtained symbolic representation and further evaluation of conversations. Intents on the agent and customer sides were defined in accord with the QA checklist provided by a Belgian call centre, where some of them express desired behaviours (e.g., welcome), and others unwished ones (e.g., doubt). Our intent recognition solution allows for classifying fragments of texts that do not necessarily match an utterance. To do this, we present a new architecture with a sliding window, that allows classifying pieces or sequences of utterances, which makes the difference from RASA and DialogFlow intent recognition. Our demo allows users to play with different window sizes.

The more precise the learning of the intents, the better results are obtained in further stages. Since in dialogues there are dependencies between utterances and the local information can change as the dialogue proceeds, we introduce a new approach for learning intents modelling the context explicitly by adding the dependence relations of intents and utterances as features. Different classifiers and textual representations were applied for studying the effect of precedences in intent

learning. The achieved results improve those obtained where intents are classified in isolation, which are also better than those obtained by RASA. Improving the F1 measure from 94% to 99% means a reduction of the error with a factor of 6. Our solution is integrated into the Nixxis Contact Suite[1].

Afterwards, our proposal is in charge of creating the symbolic representation of conversations based on their intent sequences. This representation allows us to see when an intent is discovered. Besides, the positions and the frequency of the intents in the calls are shown. The symbolic representation is sensitive to the confidence threshold values estimated for each intent. Finally, to offer an automatic evaluation of dialogues (see Figure 2), the subsets of intents corresponding to each QA checklist item were identified, and then, a rule-based system was defined to aggregate the intent confidence values. Our demo allows users to explore different intent confidence thresholds and their influence on the symbolic representation, call evaluation, explanation, and interpretation.

Figure 2: *Automatic evaluation of a conversation.*

3. Patterns, insights, and recommendations

The intent recognition module does not only serve as the basis for call evaluation but also supports recommendations, assists agents, and discovers insights. Our proposal can detect what the most frequent intents are, analyse if the agents mostly express doubt or give trust to the customers, and identify calls where the customer's problem is not solved.

We develop the dialogue description notebook, a tool for extracting general statistics about the vocabulary, times, utterances, intents, and hold times. Statistics allow users to perform various analyses, such as: speaking time partition, the average call time, the average duration of utterances, the presence of hold times, their duration, and relation to the intent "request wait" (see Figure 3). Note that a single long-duration hold time appears in the call and it is preceded by the "request wait" intent (i.e., "*je vais vous mettre en attente*").

Users can also see how more extracted patterns have potential applications in recommendations. The first implemented Recommender System (RS) was the automatic filling in of the QA checklist where we see recommendation as a classification problem. We have already discovered other patterns to support more advanced recommendations (e.g., hold times and their causes, main call reason topics, and intent behaviour). Based on those patterns an RS could be developed to suggest to an agent what to improve based on his call history. We could also make actionable advice; for instance, suggest what deficiencies the supervisor should work on agents to improve their performance, and recommend to decision-makers a better call outline.

[1] https://www.nixxis.com/productsandservices/ncs/

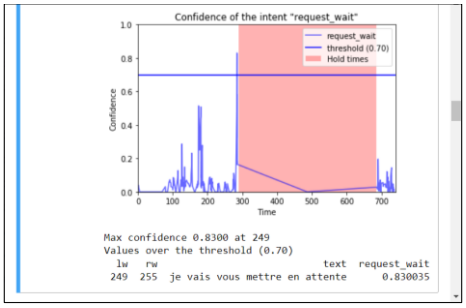

Figure 3: *The intent "request wait" and subsequent hold time.*

Recommendations and assistance could also be given during the call to suggest how to proceed in a certain situation. Given the importance of recommending sale offers as part of the conversation, we focus on this high-value use case from the business perspective. We approached the next intent online recommendation as a sequence-based RS which suggests to the speaker the optimal approach in a dialogue given the sequence of previous intents. Our demo illustrates the influence of the window size, and the binary and multiclass prediction in recommendations.

4. Conclusions

The developed system that completes a call QA checklist based on a high-level symbolic representation of calls allows moving from a few to 100% analysed conversations and detecting fine-grained reasons for QA problems. Our analysis of conversations can be used at different stages of the quality feedback loop, and support: agents, supervisors, high-level decision-makers, and clients. The defined architecture with a sliding window that allows classifying pieces or sequences of utterances and modelling the context explicitly by including the dependence relations of intents and utterances as features makes the difference from other intent recognition solutions. Besides, the main discovered patterns empower contact centres to suggest training opportunities on how an agent can improve, give an indication of which agent a supervisor should listen to, or assist agents (e.g., when proposing sales).

5. Acknowledgments

This research was carried out as part of the SYNAPS project in collaboration with Nixxis, ETRO, and a Belgian call centre, funded by Innoviris. The authors gratefully acknowledge the VUB AI applied research team members, Izmir Khalish, and Benjamin Jan Vermunicht for their contribution.

6. References

[1] B. Ar, F. Béchet, and B. Favre, "CallAn: A Tool to Analyze Call Center Conversations," in *Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS)*, 2016, pp. 1–6.

[2] S. Roy, R. Mariappan, S. Dandapat, S. Srivastava, S. Galhotra, and B. Peddamuthu, "QART: A system for real-time holistic quality assurance for contact center dialogues," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 3768–3775.

[3] N. Mehrbod, "Forecasting and controlling key performance indicators in call centers," *Res. Sq.*, 2021.