Summarizing Behaviour: An Experiment on the Annotation of Call-Centre Conversations

Morena Danieli[1], Balamurali A R[2], Evgeny A. Stepanov[1]
Benoit Favre[2], Frederic Bechet[2], Giuseppe Riccardi[1]
[1]Signals and Interactive Systems Lab, University of Trento, Trento, Italy
[2]Aix-Marseille Université, CNRS, LIF UMR 7279,
13000, Marseille, France
Corresponding author: morena.danieli@unitn.it

Abstract

Annotating and predicting behavioural aspects in conversations is becoming critical in the conversational analytics industry. In this paper we look into inter-annotator agreement of agent behaviour dimensions on two call center corpora. We find that the task can be annotated consistently over time, but that subjectivity issues impacts the quality of the annotation. The reformulation of some of the annotated dimensions is suggested in order to improve agreement.

Keywords: behavioural analytics, corpora annotation, inter-annotator agreement

1. Introduction

One of the most critical problems of contact centre business is the quality assessment of the huge amount of inbound and outbound phone calls they receive every day. Calls between customers and agents need to meet several quality requirements that include both objective measures, like first call resolution and monitoring of caller waiting time, and more intangible criteria related with the communicative appropriateness of the agents. Some contact center companies constantly evaluate the quality of the calls by applying observational methods such as qualified listening of randomly selected conversations. While those methods present some advantages, they are highly subjective, and run the risk of being statistically invalid since the traditional observational methods often target only 2 to 10 calls (or interactions) per agent per month.

In the SENSEI project we design, develop, and evaluate methods and tools based on analytics technologies that may be helpful for approaching the problem described above. The goal of SENSEI is to provide different types of summaries of huge amount of spoken and written conversations, including call center conversations and social media interactions. In this paper we focus on the evaluation of the behavioural annotation of call center conversations.

2. The Annotation Task

In a real call centre the live conversations are assessed by Quality Assurance (QA henceforth) supervisors and are scored against established contact handling criteria, summarised into a QA questionnaire. The typical survey asks questions like "Was the agent able to provide the client with the listening to the calls, the QA supervisor judges if the agent passes of fails the criterion. Most of the questions are focused on the behavioural attitude of the agent, and replying to them is a highly subjective task. Starting from the analysis of actually used Agent Conversation Observation Forms (ACOF), we developed observation surveys that were applied to two corpora of phone conversations, the Italian corpus LUNA and the French corpus RATP-DECODA. Both corpora are collection of inbound calls, but they come from different semantic domains: the LUNA conversations are calls to technical assistance services by corporate users, the DECODA ones are information inquiry calls by generic users. This difference implies different requirements for the work of call center agents, so it was necessary to define for each corpus a specific questionnaire with slightly different Quality Monitoring (QM henceforth) parameters.

The annotation task involved a group of eight QA supervisors (five for the Italian corpus, and three for the French corpus). They were instructed to familiarize with the survey forms and received information about the semantics of the scenarios. Each QA supervisor was asked to fill an ACOF for each observed call. Both the Italian and the French annotator groups worked on the same sample of conversations for each language, so that for each conversation of the sample we had multiple ACOFs.

The first step of our evaluation protocol had the aim of evaluating the stability of the ACOFs. We applied a test/retest protocol. The test-retest protocol is commonly used in experimental psychology as a method for assessing the stability of a psychological construct over time. A classic example of its application in psychophysiology research is described by McKinney *et al.* (1985). The protocol requires that the same test is given to the same subject in two separate sessions (T1 and T2).

The scores on the two occasions are then correlated to get the coefficient of stability of the test. The closer each respondent's scores are on T1 and T2, the more reliable the test measure is. A coefficient of stability of 1 says that each subject scores are perfectly correlated. That is, each

subject scored the exact same thing on T1 as they did on T2. A coefficient correlation of 0 indicates that the scores at T1 were completely unrelated to the scores at T2; therefore the test is not reliable.

For SENSEI we designed the following test-retest protocol. We recruited two annotators who contributed to the annotation of ACOFs of LUNA and RATP-DECODA conversations. Each of them received 60 conversations: half of those conversations were extracted from the ones they annotated from LUNA, half from the ones they annotated from DECODA. 34 of the selected conversation had been annotated less than 41 days before the retest, 26 had been annotated more than 41 days and less than 90 days before the retest.

The annotators worked independently, and without having access to their previous ratings. They received instructions for re-annotating each item of the ACOFs over the selected data, without worrying about the fact that they did not remember the conversations they already annotated.

We calculated the test-retest correlation by using the Intraclass Correlation Coefficient (Zou, 2012), whose formula is ICC = (F - 1) = (F + *k* - 1), where F is the F ratio, and *k* is the number of tests. We do not have missing observations in this experiment. In Table 1 we report the results of test-retest. In total, 60 conversations have been re-annotated by each participant. The scoring was calculated by counting 1 each time the annotator attributed at T2 the same value (Pass, Fail, or NotApplicable) s/he attributed at T1 for each item of the ACOF, 0 in case of difference.

T1-T2 (*n* dialogs)	**F for Subjects (confidence level %)**	**ICC and Confidence Limits**		
		ICC	***Lower***	***Upper***
< 40 days (34)	29.2 (90)	0.93	0.84	0.97
41-90 days (26)	33.8 (90)	0.80	0.66	0.82

Table 1: ICC for Test-Retest Experiment.

The test-retest experiment showed that even when T1 and T2 are in the interval 41-90, reliability is still good. This result supports the hypothesis that the ACOF may be a stable evaluation tool over time.

3. Data Sets

As it was already mentioned, the Agent Conversation and Observation Forms (ACOF) were annotated for two corpora: Italian LUNA corpus (Dinarelli *et al*. 2009) and French RATP-DECODA corpus (Bechet *et al*. 2012).

The Italian LUNA corpus is a collection of 572 dialogues in the hardware-software help desk domain. The dialogues are conversations between agents and corporate users engaged in problem solving. A subset of 300 dialogues has been used for ACOF annotation.

The French RATP-DECODA corpus [1], on the other hand, is a collection of 1,514 dialogues of a Paris public transport authority (RATP) call center. The corpus focuses on conversations between agents and public transportation customers on the topics of transportation routes, lost items, etc. A subset of 288 dialogues has been used for ACOF annotation.

The two corpora are similar in genre. However, they differ with respect to the users agents have to converse with: corporate vs. 'common' users. This difference makes these two corpora particularly interesting for behavioural annotation.

Besides ACOF annotation both corpora have been annotated for other levels of information, such as concept-attributes and values, dialogue acts, predicate-argument structures for LUNA [2]; and Named Entities and syntactic information for DECODA [1]. This non-behavioural annotation can be utilized for the studies on automatic detection of behavioural patterns in call centre conversations. Since automatic detection of behavioural patterns relies on the consistency of annotations, despite the subjective nature of the task, in the next section we report evaluation on the annotation consistency.

ID	**Quality Monitoring Parameters**
1	Agent respects opening procedure
2	Agent listens actively and asks relevant questions
3	Agent shows the information in a clear, comprehensive and essential way
4	Agent manages the objections reassuring the customer and always focusing on client satisfaction
5	Agent manages the call with safety
6	Agent uses positive words
7	Agent follows the closing script
8	Agent is polite and proactive with the customer
9	Agent is able to adapt to the style of client's communication always maintaining professionalism
10	Agent Management: he negotiates the wait always giving reasons
11	Ability to listen
12	Takes care of the problem with the next re-contact

Table 2: Evaluated Quality Monitoring Parameters

4. Evaluation of Annotation Consistency

4.1 Annotation for QM

DECODA conversation corpus is annotated for the QA form mentioned in Table 4 by near native annotators, whereas Luna conversation corpus is annotated by native speakers. The audio and respective transcript were provided to the annotators. The annotators are QA supervisors in one of the largest call centre company in Europe. Based on the specific questions mentioned in Table 4, they had to mark the conversation as PASS, FAIL and NA. The category PASS reflects that annotator is satisfied with specific objective mentioned in QM

questionnaire. If they are unsatisfactory, then they are marked as FAIL. If the annotators do not have sufficient information to make decision they are marked as NA. This includes cases in which service does not provide any actions of up-selling, or if agent does not collect specific information of the customer like name, surname etc. or if the agent's objective (qualitative or quantitative) is ambiguous.

4.2 Pre-process and Agreement Calculation

Fleiss Kappa is used as metric for inter-annotation agreement (Fleiss, 1971). It measures reliability of agreement between a fixed number of annotators when assigning categorical ratings to a number of items. Annotation was performed in a controlled setting. After the completion of the annotation procedure, we observed that there exists a high degree of disagreement on the LUNA corpus. One possible reason for this is the way the annotation procedure was carried out. The annotation was completed in batches of two. Thus to remove the differences, we calculated agreement in those batches. The kappa agreement along with data statistics of annotation are provided in Table 3 and Table 4.

4.3 Analysis of the DECODA Corpus

We observe, on an average, that there is a moderate inter-annotation agreement. In some cases, for instance questions 2, 4 and 6, it is poor or no agreement at all. Given this premise, the QA task is highly subjective. The reasons for this subjective nature are many. Human bias is an important factor in such annotation tasks. For instance, Question 6 has no agreement at all. This QM tries to find out whether the Agent uses positive words during the conversation with the customer. However, since agents do not use very negative (rude words for example) or very positive words, the association of words with sentiment labels is highly subjective. In addition, agents recorded in the DECODA corpus were not instructed to behave so that they obtain a good score with the QA questions. Another factor is the different dimension each QM parameter tries to address. Question 3 has a moderate agreement because it is multi-faceted. It tries to assess agents' behaviour on three dimensions - clarity of speech, comprehension, and finally on picking necessary information from the customer. From the perspective of QM supervisors, this question is overloaded and assessing it can be very difficult.

From the annotation statistics in Table 3, only questions 1 and 10 have high degree of agreement. These two questions could be evaluated automatically. Conversely, critical QM parameters like Question 2, 3, 4, 8 and 11 are more difficult to answer for an automatic system. Moreover, FAIL samples are too few to create a sound supervised approach for evaluate each QM parameter.

Lastly, it appears also that although annotators would agree on the fact that a conversation is problematic, they might differ on the FAIL parameters given to justify their decision. Therefore, as the agreement at the question level Is moderate, it would be difficult to design a high confidence automatic QM evaluation system which answers separately to each of them.

Q.ID	PASS	FAIL	NA	*k*	Agreement
1	286	2	0	1.0	Perfect
2	268	11	9	0.08	Slight
3	276	4	8	0.44	Moderate
4	212	16	60	0.15	Slight
5	274	11	3	0.51	Moderate
6	154	67	67	-0.13	No
7	274	3	11	0.54	Moderate
8	283	2	3	0.65	Moderate
9	278	6	4	0.53	Moderate
10	175	3	110	0.90	Perfect
11	270	12	6	0.25	Fair

Table 3: DECODA Corpus: Inter-annotation agreement using Fleiss Kappa along with category selected based on majority voting

4.4 Analysis of the LUNA Corpus

Compared to the DECODA corpus, LUNA annotation agreement is lower. Kappa measurements suggest on an average, a fair agreement. Part 2 annotation is carried out under a stricter environment. From Table 4, it is evident that Part 2 annotation has better agreement than Part 1. Hence, further discussions on the LUNA corpus only address Part 2.

Question 1, 7 and 8 have good agreement. These questions address procedural aspects and are therefore easily tangible. Compared to these, questions 2, 3, 6, 10 and 11 have slight agreement. Question 2, 3 and 10 are multi-dimensional. Evidently, the agreement is low. In the DECODA corpus too, we see the same effect. This concretely suggests dissecting the multi-dimensional QM parameters into single dimensions. For instance, Question 2 asks the supervisor whether *the Agent listens actively and asks relevant questions*. This addresses the agent's listening ability as well his/her ability to ask relevant question. To have better agreement, perhaps these aspects should be assessed individually.

Another reason for lesser agreement, in general, is the inability to quantify the parameters that are being judged. For instance, question 11 checks the ability to listen. This is highly subjective and cannot be exactly quantified. May be more quantization is required in assessing such parameters.

5. Conclusions

Evaluating agreement on annotation tasks that include subjective judges is problematic due to the subjectivity of such tasks. In this paper we report the assessment of inter-annotation agreement for an annotation task of two corpora of call centre conversations. The annotators needed to judge the call centre agents' communication behaviour based on a set of QM parameters arranged on a questionnaire inspired to the ones currently used by the QA supervisors in call centre. The results showed that while some QM parameters reported moderate to good agreement in both the corpora, the multi-dimensional

Question ID	PASS	FAIL	NA	*K*	Part 1 Agreement	*K*	Part 2 Agreement
1	13	0	287	0.31	Fair	0.66	Substantial
2	290	7	3	0.11	Slight	0.12	Slight
3	288	4	8	0.07	Slight	0.19	Slight
4	262	11	27	0.11	Slight	0.22	Fair
5	293	4	3	0.21	Fair	0.26	Fair
6	269	29	2	0.16	Slight	0.15	Slight
7	294	0	6	0.71	Substantial	0.94	Perfect
8	298	0	2	0.39	Fair	0.62	Substantial
9	294	4	2	0.30	Fair	0.37	Fair
10	12	6	282	0.24	Fair	0.07	Slight
11	297	1	2	0.50	Moderate	0.08	Slight
12	180	10	110	0.26	Fair	0.20	Fair

Table 4:
LUNA Corpus: Interannotator Agreement using Fleiss Kappa along with category selected based on majority voting

parameters are more sensitive to the high subjectivity of this annotation task. On the basis of those results we can suggest that the introduction, whenever possible, of single-dimensional parameters may reduce the subjective bias of behavioural annotation tasks.

6. Bibliographical References

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Beze, M., De Mori, R. and Arbillot, A. (2012). Decoda: a call-centre human-human spoken conversation corpus. In *Proceedings of LREC*, pp. 1343- 1347.

Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A. and Riccardi. G. (2009) Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. In *Psychological bulletin*, 76(5):378.

McKinney, Mark E., *et al*. (1985). The Standardized Mental Stress Test Protocol: Test‐Retest Reliability and Comparison with Ambulatory Blood Pressure Monitoring. In *Psychophysiology* 22.4: 453-463.

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in medicine*, 31(29), 3972-3981.

Acknowledgement

The research described in this paper is funded by the the SENSEI project (SENSEI FP7-ICT-610916) that include the following partnership: University of Trento (Coordinator), Université d'Aix Marseille, University of Sheffield, University of Essex, Teleperformance, and Websays.