Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-30

Collection of Resources and Evaluation of Customer Support Chatbot

Daiga DEKSNE and Andrejs VASIĻJEVS[1]
Tilde, Latvia

Abstract. In this study, we propose a practical approach to creating a chatbot on the basis of data accumulated by customer support operations. The selected use case is a typical representative of a chatbot application in customer service centers that want to improve their efficiency and raise customer satisfaction. We show how company support information and logs from support interactions can serve as source data for creating a customer support chatbot. The chatbot developed in this use case is targeted at Latvian-speaking users and demonstrates an implementation of Q&A functionality in the Latvian language. We also propose a simple evaluation metric for chatbot responses to natural language questions. As practical chatbots cannot be perfect in providing appropriate answers to all user questions, this metric can be used to assess the readiness of the chatbot for being released to real users. In our experiment, a chatbot with a score of 0.45 showed positive results in a user survey.

Keywords. Intelligent virtual agents, chatbots, dialog systems, chatbot evaluation

1. Introduction

As social platform technologies and innovations in artificial intelligence are rapidly developing, we are witnessing a rise of social chatbots – intelligent virtual assistants that communicate with users on social media platforms like *Facebook*, *Skype*, or *Twitter*.

In this study, we propose a practical approach to creating a chatbot on the basis of data accumulated by customer support operations. The selected use case is a typical representative of a chatbot application in customer service centers that want to improve their efficiency and raise customer satisfaction.

The chatbot developed in this use case is targeted at Latvian-speaking users. As the Latvian language is only rudimentary supported in online chatbot development platforms like *Facebook*, *wit.ai* and *Microsoft Cognitive Services*, we created a full implementation of chatbot Q&A functionality for the Latvian language. To our knowledge, this is the first work on customer service chatbots for the Latvian language.

We also present a simple and easy-to-implement approach for evaluating a chatbot's ability to answer free-form questions, as well as provide an initial analysis of usage patterns.

[1] Corresponding Author: Andrejs Vasiljevs, Tilde, Vienibas Gatve 75a, Riga, LV1004, Latvia; E-mail: andrejs@tilde.lv.

2. A Use Case of Customer Support Chatbot

This case study is based on the development of a chatbot for a telecommunication company (hereinafter – the Company) that provides Internet connectivity and related services to a large base of individual and business customers.

A large part of questions addressed to the operators of a customer service has straightforward answers. As we have noted in our previous study [1], the motivation of the Company for introducing a customer service bot is the need to decrease customer call center workload and to improve communication with clients by reducing time user is obliged to wait for a human customer support operator to become available. The benefits of using the bots in customer service – the service is not restricted to the certain working hours, it is available on 24/7 basis. The bot can handle simple, routine tasks, only more advanced tasks are forwarded to a human operator.

In our use case the chatbot was focused on the most frequent category of questions – those related to billing and payments.

3. Collection of Resources

When starting development of a new AI conversational agent for customer service, it is important to learn beforehand what questions the bot's potential conversation partners will ask. Although the bot is designed to work in a particular sphere – to provide customer service in some area – people tend to ask questions of a general nature or even try to outsmart the bot, to find its weak points.

What are the potentially valuable data sources to use? We have explored the following: the Company's Frequently Asked Questions webpage, the Company's Twitter feed, discussions on the Company's forum webpage, e-mails processed by the Company's customer service, and logs of the customer support interactions on the Company's live chat window.

3.1. FAQ and Support Information on the Company's Website

The Company's Frequently Asked Questions webpage provides good structured data, which is usually grouped by a topic. This is very important, as one of the most important tasks for a bot is to detect an intent in a customer's utterance, which helps providing an adequate answer. The bot can learn from this data applying machine learning techniques. The data, in turn, can be processed semi-automatically.

We started with extraction of triplets from this data – category, question, answer. The main deficiency of this data is that questions are formulated in precise literary language and are very formal. It is not a language in which potential customers will ask a question in a conversation.

The same applies to the answers: the language is too formal for a conversation, and the text is too long. Some simplification and summarization of questions and answers should be done manually. Manual work was also required to concretize the intent of every question. Nevertheless, this data formed the basis of the corpus used in development of the bot.

Discussions on the company's forum webpage more closely resemble a spoken language, but this data lacks a good structure. Although every discussion is grouped under some topic that could resemble a question, discussion texts are very 'noisy' – many

emotions, many contradictive points of view. Manual work was involved to review discussions on the forum, to form a question, and to map it to an intent.

3.2. Logs of Live Chat with the Call Center Operators

Logs of the Company's live chat window contain short utterances typical to spoken language. They reveal how a potential client could talk with a customer service bot. A problem for an automatic procession of such text is to detect the boundaries of context. In a single utterance, a costumer could refer to some information spoken about previously, as utterances are context dependent.

For detection of the most common utterances and topics, we have employed some statistical methods. We have sorted user utterances by frequency. The top frequency is about 4,000. The most frequent utterances are: *'good day'*, *'thank you'*, *'hello'*, *'yes'*, *'good evening'*, *'ok, thank you'*, *'thank you for your answer'*, and similar.

Phrases with a frequency above 10 reveal almost nothing about the topics that interest the costumers most. In addition, phrases with a frequency slightly below 10 do not help in this task either. They are short context-dependent answers to the previous discussion, for example: *'open-ended agreement'*, *'how much'*, *'with a courier'*, *'euro'*, *'what I have to do'*. Our conclusion: the way in which customer can formulate a question is very diverse.

However, there is another aspect in this data from which the bot can learn. We have taught the bot to answer to the different ways customers express greetings, the gratitude about answers given by the customer service representative, and agreement or disagreement with something.

By sorting and analyzing tweets, we have learned how users formulate questions, for example: '*is it possible …'*, *'where can I …'*, *'how can I …'*, *'do you have …'*, and etc.

3.3. Client Support E-mails

The language of e-mails received by customer service is rather formal. There is not much to be done with an automatic procession. We did some manual work to make a list of the typical problems and to assign the corresponding intents.

As we concluded while analyzing the frequency of user utterances, the formulation of a problem can be very diverse. We have applied some simple methods for paraphrasing the question to cover a potentially larger amount of potential user utterances. We duplicate phrases by paraphrasing certain verbs in infinitive verb phrases. For example, we make paraphrases with *'I want to …'*, *'I need to ..'*, *'I would like to …'*, *'is it possible to …'*, and etc.

3.4. Resources Resulting from Data Processing

This section provides summary statistics of collected data and the resources that resulted from processing this data.

The following data was collected for analysis and processing:

- 117,235 utterances from live chat logs
- 890 tweets from Twitter logs
- 933 e-mails received by customer support

Using semi-automated and manual process, we extracted from this data 758 user questions related to the billing – the primary scenario defined for the chatbot.

We enriched this collection by verb paraphrasing as described in Section 3.3. This more than tripled the resource to 2,319 questions.

To enrich functionality of the chatbot, we also used the collected source data to extract user questions about other topics besides billing. This yielded 2,720 questions. After paraphrasing of verbal phrases, we got 12,293 variations of user questions on different topics.

All these questions were manually mapped to the 250 user intents. For every intent, we created one or several custom answers to be provided by the chatbot. To diversify a dialogue, the chatbot chooses the answer randomly, if there are several available. These answers are based on the answers provided by the call center operators. We rewrote these answers to ensure that they are consistent, informative, short, and in the conversational genre.

To avoid lengthy answers, we often included links to the respective support information on the Company's website.

As dialog processing in our system is based on AIML technology (Artificial Intelligence Markup Language; [2]), we automatically generated 38,314 AIML patterns from the collected and paraphrased questions (14,612 in total), allowing unknown words before and/or after a phrase. Table 1 shows some examples of such AIML patterns and the corresponding phrases, for which the same intent will be assigned. Template tags (*<template></template>*) contain intent identification.

Table 1. AIML samples and corresponding phrases.

AIML sample	Phrase

`<category>`
`<pattern>SAMAKSĀT MĒNEŠA BEIGĀS</pattern>`
`<template>{rekins_parnest_maksu}</template>`
`</category>`
	to pay at the end of the month

`<category>`
`<pattern>* SAMAKSĀT MĒNEŠA BEIGĀS</pattern>`
`<template>{rekins_parnest_maksu}</template>`
`</category>`
	is it possible to pay at the end of the month'

`<category>`
`<pattern>SAMAKSĀT MĒNEŠA BEIGĀS *</pattern>`
`<template>{rekins_parnest_maksu}</template>`
`</category>`
	to pay at the end of the month, is it possible?

`<category>`
`<pattern>* SAMAKSĀT MĒNEŠA BEIGĀS *</pattern>`
`<template>{rekins_parnest_maksu}</template>`
`</category>`
	I wish to pay at the end of the month, is it possible?

We added additional 550 AIML patterns to cover generic conversational expressions like greetings, gratitude expressions, agreement, disagreement, small talk, etc. For example, to the user's utterance *'I have a question'* bot will answer *'I'll try to answer'*. The lists of such patterns where compiled by a linguist.

4. Evaluation

Radziwill & Benton [3] review various approaches in chatbot quality assessment, aligning them with ISO 9241 concept of usability: "The effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" [4]. They list numerous quality attributes to measure efficiency, effectiveness, and satisfaction. In a practical setup, assessment of all these attributes is not always feasible due to time and resource restrictions. In our work, we adapted a simplified approach that allows quickly assessing responses of the chatbot and its readiness for introducing to the real users.

4.1. Evaluation of the Natural Language Q&A

Shawar & Atwell [5] label responses of a chatbot as *reasonable reply, weird but understandable*, or *nonsensical reply.* Yu et al. [6] use similar annotation schema to label answers of the generic chatbot – *Appropriate*, *Interpretable* and *Inappropriate*.

To evaluate answers of the customer support chatbot, we modified and extended this labeling schema to five labels – *Appropriate, Partly appropriate, Not knowing, Partly inappropriate, Inappropriate* – as shown in Table 2. In our schema, the label *Interpretable* is replaced with *Not knowing.* This label is for quite typical answers in situations where the chatbot is not able to link user question to any of the known intent, and the option to connect with a support specialist – human person – is offered to the user.

Table 2. Labels and their scores for annotation of the chatbot's answers.

Label	Definition	Answer to the question 'How to reset my device?'	Score
Appropriate	Coherent with the user utterance, providing concrete answer to the user's question	Chatbot: *Press the button on the left side of the device and keep it pressed for 5 seconds.*	1
Partly appropriate	Generic response that only partly addresses user's question or responds to the user's utterance	Chatbot: *For solutions to technical issues please check the online manual <link>*	0.5
Not knowing	Neutral response that does not answer the question and does not confuse the user	Chatbot: *There are some questions that I am not able to answer yet. Would you like to be connected with a human support specialist?*	0
Partly inappropriate	Response has only distant relation to the user's utterance and may be confusing to the user	Chatbot: *If Internet connection is lost try to reset your device*	-1
Inappropriate	Not coherent with the user's utterance and utterly confusing to the user	Chatbot: *Your password should be at least 8 characters long*	-2

We attributed respective numeric scores 1, 0.5, 0, -1, -2 to these labels. Positive and negative scores are asymmetrical, as our initial studies suggested that inappropriate answers have a stronger negative effect on user perception than appropriate answers. Users are mostly tolerant to limitations in the chatbot's competence that are exposed by

answers in the category *Not knowing* and *Partly appropriate*; but they badly tolerate inappropriate answers that are completely out of the context of the user's utterance.

We used a test set of 100 user questions and performed human assessment using our labeling schema. The initial test set was based on questions from live chat logs with human operators at the call center. The data used for composition of the test set did not overlap with the data used for chatbot's development.

We calculated the average quality score of the chatbot by dividing the total sum of scores with the number of questions. This score can range from 1 (ideal) to -2 (fully inappropriate).

During chatbot's development, we performed periodic evaluations to assess quality improvements. The initial score of the first evaluation was -0.65, indicating a poor ability to provide an appropriate response. After several iterations of improvements, we achieved a score of 0.45. This version was released to beta testing and a user survey, as described in the next section.

4.2. User Survey

A beta test was performed before publishing the chatbot in production. 79 users were invited to participate, of which 48 responded and filled in the survey form about their experience. 65% of the survey participants were female and 35% male. For 63.3% of the beta users this was the first experience chatting with a bot.

The survey questions included time spent on chatting, perception of the chatbot's personality, ease of use of the guided dialog, and overall assessment of the chatting experience. The free-form natural language questions asked by users were assessed by analyzing the usage log as described in Section 5.

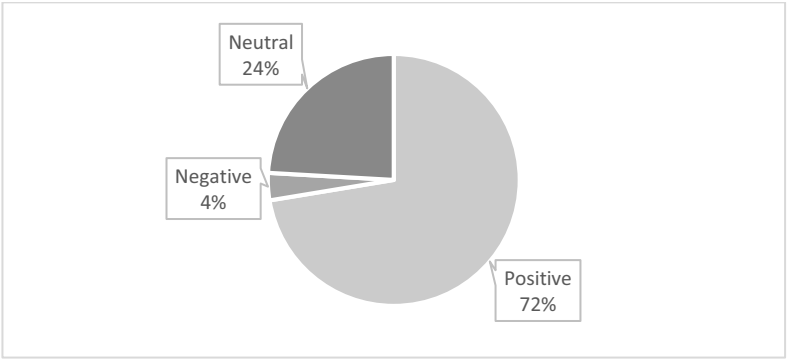

Figure 1. User response to the question about the overall experience of chatting with the customer support chatbot.

60% spent more than 10 minutes chatting with the bot; 40% spent from 5 to 10 minutes, and no one spent less than 5 minutes. 76% liked that the chatbot has a female personality. 21% would prefer to have a robot personality, and only 1 user would rather want a male personality for the client support chatbot. It seems that user gender does not play a significant role in the preference of the chatbot's gender. This confirms with the pattern seen in other studies [7].

Assessing the guided dialog to solve the most typical technical issues, 79% of users answered that it was easy to understand and helpful. Other users were less satisfied and pointed to specific elements of the guided dialog. Only two users reported dissatisfaction due to problems getting the answer they are looking for from the guided dialog. This confirms that a guided dialog may be a successful solution for the typical questions.

Despite the fact the chatbot was not able to provide appropriate answers to all the questions, 72% of users answered that they had a positive experience using it; 24% were neutral; and only one user reported negative experience from chatting with the chatbot (see Figure 1).

5. Analysis of Usage

We also analyzed logs of chatbot usage in the beta testing period. This helps to identify several usage trends that should be taken into account for improving user experience. Although testers were instructed that the chatbot is specialized for questions related to billing and the most typical technical questions, 45% of questions were completely unrelated to the Company's services, for example *'How old are you?'*, *'Where do you live?'*, etc.

The major challenge is to deal with ungrammatical language in user input – spelling errors, missing diacritics (there are 11 characters with diacritics in the Latvian alphabet), slang, wrong syntax, etc. 23% of utterances showed one or several such problems. Current implementation of our chatbot has a limited ability to process some simple typical errors.

6. Conclusion

In our study, we have shown how company support information and logs from support interactions with customers can serve as a source data for creating a customer support chatbot. We have described how this data can be used to generate AIML patterns.

We have also proposed a simple evaluation metric for chatbot responses to natural language questions.

As practical chatbots cannot be perfect in providing appropriate answers to all user questions, this metric can be used to assess the readiness of the chatbot for being released to the real users. In our experiment, a chatbot with the score 0.45 showed a positive result in the user survey.

To our knowledge, this is the first work on customer service chatbots for Latvian language.

Acknowledgments

The research leading to these results has received funding from the research project "Competence Centre of Information and Communication Technologies" of EU Structural funds, contract No. 1.2.1.1/16/A/007 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 2.4 "Multi-modal human-computer interaction in multilingual environment".

References

[1] A. Vasiljevs, I. Skadina, D. Deksne, M. Kalis, I. Vira, Application of Virtual Agents for Delivery of Information Services. *New Challenges of Economic and Business Development* (2017). 667-678.
[2] M. G. B. Marietto, R. V. Aguiar, G. O. Barbosa, W. T. Botelho, E. Pimentel, R. S. França, V. L. Silva, Artificial Intelligence Markup Language: A Brief Tutorial. *International Journal of Computer science and engineering Survey (IJCSES)*, Vol. 4, No. 3, June 2013.
[3] N. M. Radziwill, M. C. Benton, Evaluating Quality of Chatbots and Intelligent Conversational Agents. *arXiv preprint arXiv:1704.04579* (2017).
[4] A. Abran, A. Khelifi, W. Suryn, A. Seffah, Consolidating the ISO usability models. In *Proceedings of 11th international software quality management conference* (Vol. 2003). 23-25.
[5] B. A. Shawar, E. Atwell, Different measurements metrics to evaluate a chatbot system. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies* NAACL-HLT 2017, Association for Computational Linguistics (2017). 89-96.
[6] Z. Yu, Z. Xu, A. W. Black, A. Rudnicky, Chatbot evaluation and database expansion via crowdsourcing. In *Proceedings of the RE-WOCHAT workshop of LREC*, Portoroz, Slovenia (2016).
[7] K. Kuligowska, Commercial Chatbot: Performance Evaluation, Usability Metrics and Quality Standards of Embodied Conversational Agents. *Professionals Center for Business Research*, Vol 2 (February 2015).