





Human-in-the-Loop Conversation Agent for Customer Service

Pēteris Paikens^(✉), Artūrs Znotiņš, and Guntis Bārzdīnš

University of Latvia Institute of Mathematics and Computer Science, Riga, Latvia
{peteris.paikens,arturs.znotins,guntis.barzdins}@lumi.lv

Abstract. This paper describes a prototype system for partial automation of customer service operations of a mobile telecommunications operator with a human-in-the loop conversational agent. The agent consists of an intent detection system for identifying the types of customer requests that it can handle appropriately, a slot filling information extraction system that integrates with the customer service database for a rule-based treatment of the common scenarios, and a template-based language generation system that builds response candidates that can be approved or amended by customer service operators. The main focus of this paper is on the system architecture and machine learning system structure design, and the observations of a limited pilot study performed to evaluate the proposed system on customer messages in Latvian. We also discuss the business requirements and practical application limitations and their influence on the design of the natural language processing components.

Keywords: Conversational agents · Intent detection · NER

1 Problem Description

The use of chatbots has been growing not only in consumer applications, but is also gaining traction in attempts aim to add conversational agents as another alternative channel for customer service communications, which is a significant expense for many companies and has potential for automation.

However, as chatbots improve towards fluent and varied language, there is an ‘uncanny valley’ effect where the observed language skills give rise to expectations of true competency in solving the customers’ problems which often can not be met by the chatbots at this point, leading to customer dissatisfaction.

In this situation we proposed an approach for integrating conversation agents in the current customer service workflow, reducing operator workload. The customer service agent would be in full control over the customer communication, but the conversation can be automated for many routine cases where the customer service agent would be expected to follow standard guidelines. The notable difference from a full-scope conversational agent is the fact that covering unusual scenarios is not required as long as the agent is capable to identify when the customer is asking something that the automated agent can not understand or answer and human involvement is necessary.

2 Related Work

Published research relevant to goal-oriented conversational agents in Latvian is limited. There has been previous work on the chatbot “Anete” [16] for telecommunications provider Lattelecom, however, the technical details have not been published. There are proof of concept systems developed for customer service at an airline and the public library network [15], and there is published work on intent detection models [1] including a review of their applicability for Latvian.

There is substantial relevant related work on such agents for English and other major languages [4, 5, 12]. A major focus of recent research is work on end-to-end neural systems [13, 14, 17, 19], however, the human-in-the-loop approach requires a natural language understanding system instead of a ‘black-box’ end-to-end solution. The key natural language processing tasks of such a system are intent detection, entity recognition and information extraction, in particular ‘slot-filling’. For intent detection and slot filling tasks state of art results have been achieved with neural network approaches, mostly with recurrent neural networks and attention mechanisms [7, 11, 18]. Our earlier research [8, 21] and other teams [1, 2, 10] also support the effectiveness of neural network models for specifics of Latvian language in other NLP tasks.

The technical aspects of building human-in-the-loop conversational agent systems have not been well described in existing literature. The core concept of human-in-the loop conversational agents is not novel, we are aware of some research of such systems [6], but most known applications of this approach are proprietary, and the inner workings of these systems are not published.

3 System Architecture

The proposed system architecture, illustrated in Fig. 1, involves an intent detection system for identifying the types of customer requests that it can handle appropriately, a slot filling information extraction system that integrates with the customer service database for a rule-based treatment of the common scenarios, and a template-based language generation system that builds response candidates that can be approved or amended by customer service operators.

The operator actions in correcting the selected intent and the appropriate response continuously provide the system with new, recent training data, and the intent detection modules are periodically retrained on it.

The prototype system was developed using the Tensorflow framework in Python, and deployed as a Docker container.

4 Named Entity Recognition

The named entity recognition system is designed to identify not only common named entities such as people and organization names, but also the specific entities which would be candidates for the slot filling task such as invoice numbers,

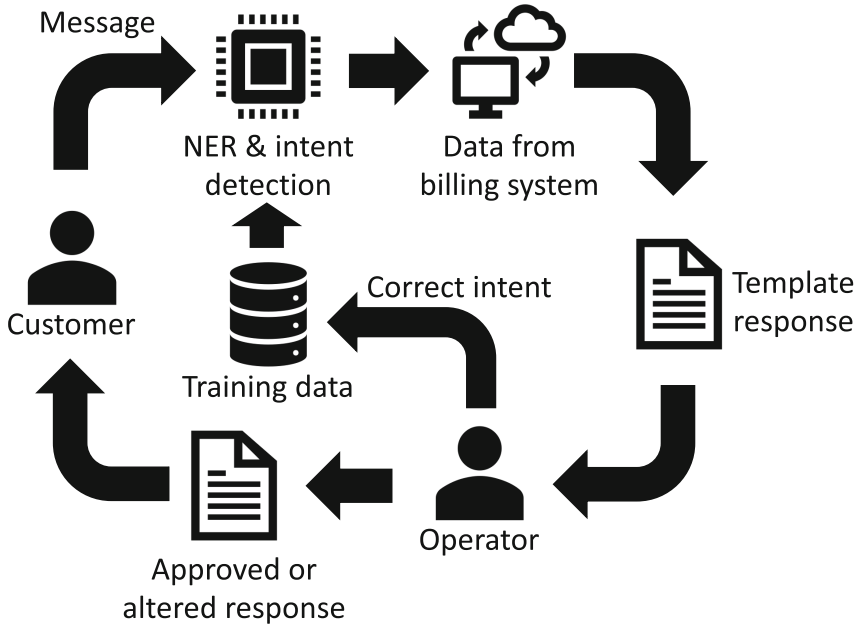


Fig. 1. System architecture

money amounts, dates, relative dates (e.g. ‘next month’) and date ranges. In total, 16 named entity categories are considered.

The dataset used for initial validation consisted of 1732 customer requests that were prepared in three steps:

1. Replace sensitive text spans with sensitive data markers
2. Manually annotate named entity spans
3. Generate named entities in place of sensitive data markers

For named entity generation, we used list of person names, registry of addresses and regular expression rules to generate invoice, personal legal ID and phone numbers.

The named entity recognition system uses GloVe word embeddings [9] pre-trained on the comment corpus collected from the project Virtual Aggression Barometer¹, character based LSTM representation, two bidirectional LSTM layers and a conditional random field (CRF) loss.

Customer requests usually contain grammatical errors and additional whitespaces for formatting. Sentences are not easily automatically separable, so full request text is used as input to maximally preserve context. Text is split on whitespace and punctuation characters without trying to extract email, date or phone number tokens. Whitespace information is passed as an additional input

¹ <http://barometers.korpuss.lv/>.

to the neural network in a one-hot vector: no space before the token, newline before the token, whitespace before the token. Word shape feature is used as an additional input to capture emails and named entities with a regular structure.

Table 1. Named entity recognition system results.

System	Precision	Recall	F1
Baseline	76.82	80.20	78.48
GloVe	80.88	78.00	79.41
BERT	81.80	80.45	81.12

Named entity recognition experiment results are shown in Table 1. The baseline system does not include any additional features. The BERT system that uses multilingual BERT model fine-tuned on the Barometer comment corpus achieves best results. The most problematic categories with F1-score below 80% are company names, product names and addresses.

As customer service discussions frequently include sensitive personal information, we implement the data minimization principle required by the General Data Protection Directive by anonymizing the customer messages using the NER results both in the intent detection system processing and in the stored training data. The customer identifying data is passed only to the main operations system, but for intent detection and permanently stored training data we replace it with generic placeholders reflecting the entity type - for example, ‘[[Phone number]’ or ‘[[Address]]’. This also has a beneficial effect on the intent detection system, reducing data sparsity, overfitting and assigning ‘superstitious’ significance to irrelevant or potentially discriminatory factors such as particular surnames.

5 Intent Detection

The intent detection system is a LSTM based deep neural network classifier. The classifier was designed to output both a coarse-grained intent topic, suitable for clustering customer requests and assigning some topics to specialized operators, and also fine-grained intent that can be matched to specific actions and answer templates. For initial word embedding layer we used GloVe [9] embeddings pre-trained on a large corpus of Latvian [20].

The developed neural network structure and chosen parameter values were the following:

- Tokenization
- Pretrained word embeddings for each token, concatenated with 10-neuron trainable ‘miniembeddings’
- Unidirectional LSTM layer with 150 cells
- Dense fully connected layer with 100 neurons, 30% dropout

- Dense fully connected layer with 50 neurons, 30% dropout
- Two separate output layers, for topics and fine-grained intent

We also investigated the application of more complex architectures such as BERT [3] which have achieved state of art results for other tasks, but this did not result in improved accuracy in our testing (see Sect. 5.1) and substantially increased training time, so this avenue was not pursued further.

5.1 Dataset and Experimental Validation

The dataset used for initial validation consisted of 1732 customer requests annotated with fine-grained intent data and named entities relevant to the intent. The data contained 24 topic classes with 115 different specific intents annotated. The intent distribution was representative of incoming customer requests, and thus was not balanced with respect to the topics. The most frequent topic class was billing with 794 requests (46% of total), which also contained the most frequent intents - postponing bills (356 requests) and confirming that an overdue bill has been paid (207 requests), while many specific intents had only a single example request.

Repeated experimentation on various options for neural network structures was performed on this dataset using cross-validation, in order to prepare a single architecture to be evaluated during the pilot study.

Table 2. Intent detection system accuracy

System	Topics	All intents	Postponing	Confirmations
Simple	68%	56%	86%	42%
Proposed	80%	70%	90%	81%
BERT	81%	69%	91%	81%

The key metrics used in evaluation (shown in Table 2) were the system average accuracy scores respectively for all the coarse grained topics, all the fine grained intents, and the F1 scores for the above-mentioned two most frequent intent groups, as they would be the focus of subsequent pilot study. The described systems include a simple multilayer perceptron without precomputed embeddings; the proposed network structure described in the previous section, and a transformer architecture based on fine-tuning BERT [3] for Latvian.

The preliminary results indicated that there was sufficient training data for two most common specific intents, and for the other topics only the coarse-grained topic classes have sufficient accuracy to be practically usable unless significant amounts of additional training data are used.

6 Slot Filling and Pre-filling a Response

If the detected customer intent is one of the prepared scenarios which can be handled by the system, then it is possible to prepare a template answer based on the detected intent and specific conditions. For example, if the intent is to change the payment plan, then it is possible to automatically verify in the core billing system whether the customer is eligible for this plan and prepare an appropriate personalized response template depending on the eligibility.

In addition, a checklist of specific actions for the customer service operator would be generated. For example, if the intent is to assert that a bill has been paid by supplying an attached payment document, then the operator needs to verify the suitability of that document.

For some intents, the system needs to extract specific information from the message in order to fulfil that intent. For example, if the customer is disputing a bill payment, then the date and amount of the payment needs to be identified. If the slot filling system in the proposed architecture would not be able to identify some of the required information, then the generated template answer would include specific sentences explicitly asking for that particular information.

This functionality would require substantial integration work with core billing systems. The proposed architecture is aimed to support this functionality, but development of it was started only after the evaluation of the pilot study and is not complete.

7 Pilot Study

The proposed model was initially validated in a three month pilot study at the mobile telecommunications operator customer service center. The pilot study was aimed to evaluate the feasibility of core technical concepts and proposed architecture in order to justify further integration and development of the full system. While the study involved the actual customer service team, it was primarily a technical feasibility pilot study without a systematic review of the human experience factors.

For the purposes of this study, the intent detection and response generation were limited to two most common types of communication - requests to postpone bill payment, and requests to restore service after payment of overdue bills.

In the scope of this pilot study, the following components of the proposed architecture were prepared and evaluated:

- Integration with message sources
- Named entity recognition
- Data anonymization
- Intent detection
- Integration with customer service systems
- Basic response templates
- Automatic retraining based on customer service agent feedback

Development of the slot filling and decision making component, as well as further work on response generation was not included in the pilot study. The pilot study

was implemented only for conversations in Latvian language, but the planned system architecture is trilingual Latvian-Russian-English. Nonetheless we believe that the scope of the pilot study is sufficient to demonstrate applicability of the full proposal.

In the limited pilot study, 14000 customer requests were processed using this system, and continuously used to retrain the intent detection model with additional data. As expected based on the preliminary testing, the detected intent and the automatically chosen answer template (which was selected for the two most frequent topics only) was accurate approximately 90% of the time and required operator intervention for the remaining 10% cases. At the end of the pilot study, the additional data gathered was able to improve the intent detection accuracy by approximately 2% points, so only 8% of the main billing requests needed changes by the operator.

From the perspective of the end users, the pilot study was considered successful, saving time and effort. From the business perspective the study affirmed the feasibility of this concept and supported continuing further development of the proposed system.

8 Conclusions and Future Work

We have described an architecture proposal for a human-in-the-loop system that supports customer service answers to customer enquiries. The initial experiments and a limited pilot study have demonstrated the feasibility of this proposal and support further development of this proposal.

It can be concluded that human-in-the-loop conversational agents are a feasible option for partial customer service business process automation. We argue (but do not conclusively prove in this study) that this approach can save time and effort when handling common customer service enquiries while still maintaining a high quality of service.

Acknowledgements. This research is funded by the Latvian Council of Science, project “Latvian Language Understanding and Generation in Human-Computer Interaction”, project No. LZP-2018/2-0216.

References

1. Balodis, K., Deksnē, D.: Fasttext-based intent detection for inflected languages. *Information* **10**(5), 161 (2019)
2. Deksnē, D.: Bidirectional LSTM tagger for latvian grammatical error detection. In: Ekštein, K. (ed.) *Text, Speech, and Dialogue*, pp. 58–68. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-27947-9_5
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding (2018)
4. Følstad, A., Nordheim, C.B., Bjørkli, C.A.: What makes users trust a chatbot for customer service? An exploratory interview study. In: Bodrunova, S.S. (ed.) *Internet Science*, pp. 194–208. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01437-7_16

5. Jenkins, M.-C., Churchill, R., Cox, S., Smith, D.: Analysis of user interaction with service oriented chatbot systems. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 76–83. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73110-8_9
6. Kucherbaev, P., Bozzon, A., Houben, G.J.: Human-aided bots. *IEEE Internet Comput.* **22**(6), 36–43 (2018)
7. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling (2016)
8. Paikens, P.: Deep neural learning approaches for Latvian morphological tagging. In: *Human Language Technologies - The Baltic Perspective*, vol. 289. IOS Press (2016). <https://doi.org/10.3233/978-1-61499-701-6-160>, <http://ebooks.iospress.nl/volumearticle/45531>
9. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
10. Pinnis, M.: Latvian tweet corpus and investigation of sentiment analysis for Latvian. In: *Baltic HLT*, pp. 112–119 (2018)
11. Qin, L., Che, W., Li, Y., Wen, H., Liu, T.: A stack-propagation framework with token-level intent detection for spoken language understanding (2019)
12. Rizk, Y., et al.: A unified conversational assistant framework for business process automation (2020)
13. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: *Thirtieth AAAI Conference on Artificial Intelligence* (2016)
14. Shah, P., Hakkani-Tur, D., Liu, B., Tur, G.: Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pp. 41–51 (2018)
15. Vasiljevs, A., Skadina, I., Deksnē, D., Martins Kalis, T., Vira, L.: Application of virtual agents for delivery of information services. In: *New Challenges of Economic and Business Development*, pp. 702–713 (2017)
16. Vevers, J.: Lattelecom klientu apkalpošanas robotmeitenes anetes projekts: soli pa solim. *Dienas Bizness* (2017). <https://www.db.lv/zinas/lattelecom-klientu-apkalposanas-robotmeitenes-anetes-projekts-soli-pa-solim-468150>
17. Vinyals, O., Le, Q.: A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015)
18. Wang, Y., Shen, Y., Jin, H.: A bi-model based RNN semantic frame parsing model for intent detection and slot filling (2018)
19. Zhong, P., Wang, D., Miao, C.: An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7492–7500 (2019)
20. Znotins, A.: Word embeddings for Latvian natural language processing tools. In: *Human Language Technologies - The Baltic Perspective*, vol. 289. IOS Press (2016). <https://doi.org/10.3233/978-1-61499-701-6-167>, <http://ebooks.iospress.nl/volumearticle/45532>
21. Znotins, A., Cirule, E.: NLP-PIPE: Latvian NLP tool pipeline. In: *Human Language Technologies - The Baltic Perspective*, vol. 307, pp. 183–189. IOS Press (2018). <https://doi.org/10.3233/978-1-61499-912-6-183>, <http://ebooks.iospress.nl/volumearticle/50320>