Chatbots with Deep Learning Techniques: the Influence of Dataset Variation on User Experience

Deola Simone[1,2][0000−0001−5531−6684], Ricardo A. Matamoros A.[1,2][0000−0002−1957−2530], and Francesco Epifania[2]

[1] Department of Computer science, University of Milano Bicocca, Milan, Italy
{s.deola1, r.matamorosaragon,}@campus.unimib.it
[2] Social Things srl, Milan, Italy
{simone.deola, ricardo.matamoros,francesco.epifania}@socialthingum.com

Abstract. The main purpose of this research is to find a correlation between the changes on chatbot training dataset and the impact on the performances of the chatbot itself.

Keywords: Chatbot · Convertion system · Deep Learning · Human Computer Interaction.

1 background of the problem

Nowadays the chatbot technology is widely used in human machine interaction[1], allowing people to interact with complex backend systems using the natural language. There are a lot of examples in the consumer market (e.g. Alexa, Siri, ok Google ecc)[2] as well as in the corporate market (e.g. chatbot for customer service). In both cases the general goal of a chatbot is to communicate properly with the user, by responding to the user request in a coherent and in the most "natural" way. Therefore, chatbots need to classify each users request according to the specific use cases they are able to manage. The chatbot classifier is trained by exploiting a set of sample phrases, taken from previous usage or handmade, divided into groups called intents. The more these phrases are similar to the phrases that will be inputted by the user, the more the classifier of the chatbot will perform well. Another characteristic that will affect the performances is the way these phrases are grouped into intents. Each intent must contain phrases that share a topic, in order to provide the chatbot pipeline with indications on how to handle each single request. In other words, the intent must contain phrases that need to be answered in a similar way by the chatbot. Due to the nature of the user interaction with a chatbot, the first setup of the dataset can become outdated during the lifetime of the chatbot. This lack of updated settings impact on the chatbot performances. This means that it is necessary to operate some changes on the chatbot to adapt it to new configurations. In order to maintain a healthy classificator, the possible changes that can be done are related to the phrases used to train it, so by removing or adding training phrases,

or by changing the intents composition, so by changing the intent assigned to each phrase, always guided by the needs of the users. For instance, there can be users asking for new information that was not included in the previous training. In this case a new intent must be added to the chatbot[3]. Other problems can be related to the responses of the chatbot, that can be vague as regards a subset of phrases belonging to some intents. In this case the intent can be split into multiple intents that cover the particular cases[4]. There also can be topics no more relevant for the chatbot so the relative intents, and relative phrases, can be eliminated.

2 problem description

Nowadays, the only way to estimate the impact of the chatbot's intents changes on its performances is done by training the chatbot itself and testing it. Those procedures are expensive as regards both time and economic costs. The purpose of this research is to train a model that can predict the impact of intents modification on the chatbot performances without the need of training it. The chatbot modifications considered will be intent removal, addition or modification. The intent modifications will be composed of phrases addition, rimotion or modification. During the data collection phase multiple chatbots will be trained, starting from the same set of phrases. The differences from one training to another will be the assignment of each phrase to an intent and the phrases selected. These changes will be done in order to simulate the possible configurations that the intent modification proposed could generate. For instance, the chatbot could be trained starting from all the phrases and then removing one group of phrases belonging to an intent at a time. The same could be done for each modification proposed. Every time a chatbot is trained, the relative performances will be calculated and assigned to the training dataset configuration. In this way, a new dataset is created, composed by the input training dataset, paired with the performances of chatbot training on it. From now on we will refer to this dataset as Performances dataset. Since the purpose of this research is to study the impact of these changes on the chatbot system in the most complete way, different chatbots and parts of the chatbot will be trained. Since the intents compositions impact mostly on the classifier part of the chatbot, the first step will be done on that. This means that the chatbot will be composed of a single NLP classifier[5], and the relative performances will be precision, recall and f1-score. Different NLP classifiers will be explored. A second step will be focused on using a complete chatbot for training and evaluation. This second test will study the impact of the dataset modification on the overall system of a chatbot. In order to evaluate the performances of these complex systems different metrics can be used, for example Dialogue efficency in terms of matching type, Dialogue quality metrics based on response type and Users' satisfaction assessment based on an open-ended request for feedback[6]. Each one of those will highlight different aspects of the chatbot. Once the Performances dataset is created, the main methods of prediction will be applied, such as regression, Neural Network, ecc,

in order to infer the performance of a chatbot[7]. During all the steps described above, in order to operate on the data, all the data manipulation techniques for Natural Language will be used (Embedding, PCA, Lemmatization, ecc.) This process will be performed on multiple chatbot's intents datasets, provided by the company. The chatbot datasets are composed of Italian phrases and cover different areas of content (banking chatbot, FAQ chatbot for different Client services, ecc..).

3 conclusions

The expected result is a model that, starting from any chatbot datasets configuration, could predict the performances of the chatbot itself, without the needs of training it. The resulting model should be computationallyl simpler than the original chatbot and should be accurate enough, in order to remove the need of the training phase for evaluation. This kind of model could help the process of creation and maintenance of a chatbot, by providing the users with an instrument that facilitates the assignment of intents to the training phrases and suggesting which phrases will be useful in the training process. A possible future implementation is to use the model to automatically correct the initial intent assignment and provide the best dataset for chatbot training, saving time and computational costs.

References

1. Ciechanowski, Leon and Przegalinska, Aleksandra and Magnuski, Mikolaj and Gloor, Peter: In the shades of the uncanny valley: An experimental study of human–chatbot interaction. Future Generation Computer Systems (2019)
2. Batish, Rachel: Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger: An experimental study of human–chatbot interaction.Packt Publishing Ltd (2018)
3. W. Ye and Q. Li: "Open Questions for Next Generation Chatbots,":2020 IEEE/ACM Symposium on Edge Computing (SEC), 2020, pp. 346-351, doi: 10.1109/SEC50012.2020.00050.
4. Richard Csaky: Deep Learning Based Chatbot Models. CoRR, dblp computer science bibliography, https://dblp.org (2019)
5. Howard, Jeremy and Ruder, Sebastian : Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
6. Shawar, Bayan Abu and Atwell, Eric: Different measurement metrics to evaluate a chatbot system. Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies (2017).
7. lmansor, Ebtesam H. and Hussain, Farookh Khadeer. Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions: Complex, Intelligent, and Software Intensive Systems, Springer International Publishing (2019).
8. Stefan Larson and Anish Mahendran and Joseph J. Peper and Christopher Clarke and Andrew Lee and Parker Hill and Jonathan K. Kummerfeld and Kevin Leach and Michael A. Laurenzano and Lingjia Tang and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction: https://dblp.org/rec/journals/corr/abs-1909-02027.bib (2019).