dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-73

Evaluation of Chatbot Prototypes for Taking the Virtual Patient's History

Andreas REISWICH[a,1], Martin HAAG[a]
[a]*GECKO Institute, Heilbronn University of Applied Sciences, Heilbronn, Germany*

Abstract. In medical education Virtual Patients (VP) are often applied to train students in different scenarios such as recording the patient's medical history or deciding a treatment option. Usually, such interactions are predefined by software logic and databases following strict rules. At this point, Natural Language Processing/Machine Learning (NLP/ML) algorithms could help to increase the overall flexibility, since most of the rules can derive directly from training data. This would allow a more sophisticated and individual conversation between student and VP. One type of technology that is heavily based on such algorithmic advances are chatbots or conversational agents. Therefore, a literature review is carried out to give insight into existing educational ideas with such agents. Besides, different prototypes are implemented for the scenario of taking the patient's medical history, responding with the classified intent of a generic anamnestic question. Although the small number of questions (n=109) leads to a high SD during evaluation, all scores (recall, precision, f1) reach already a level above 80% (micro-averaged). This shows a first promising step to use these prototypes for taking the medical history of a VP.

Keywords. natural language processing, machine learning, medical education, algorithms

1. Introduction

In the publication by Riemer and Abendroth [1], different approaches are presented how Virtual Patients (VP) can best be used in medical education. The listed systems are thereby CASUS, CAMPUS and INMEDEA [1]. Typical task types in such case-based learning systems are, for example, multiple-choice, long menu, free text, or assignment questions [2]. In the current CAMPUS system, such an interaction element is used during the anamnesis interview. Thereby, the students are able to select an appropriate anamnestic question and receive the VP answer from CAMPUS. This dialog is based upon predefined anamnestic questions, which are stored in a database. In order to build up additional competencies among the students a more flexible and individual conversation should be considered. This requires therefore a change of the system, since storing an anamnestic question in each variation isn't feasible. Hence, this research paper examines various methods from the fields of Natural Language Processing/Machine Learning (NLP/ML) to provide the capability of dealing with unknown questions. To this end, different approaches (prototypes) are implemented using Python, evaluated by leave-one-out cross-validation (LOO) and compared with each other using different

[1] Corresponding Author: Andreas Reiswich, Heilbronn University, Max-Planck-Straße 39, 74081 Heilbronn, Germany, E-Mail: andreas.reiswich@hs-heilbronn.de

micro-averaged ML scores (recall, precision, f1). The overall aim is to determine the best performing prototype, when considering a small number of given training data (n=109).

2. Methods

2.1. Literature Review

The literature databases MEDLINE[1], IEEE Digital Library[2] and ACM Digital Library[3] were used to conduct a systematic review of chatbots that were utilized in an educational environment. The underlying method was derived from the PRISMA flow diagram [3]. The year range was set for all databases from 2015 to 2018. In case of IEEE the options *Full Text & Metadata* and *My Subscribed Content* were selected for the search. For ACM, the option *ACM Full-Text Collection* and *Any field* for search terms and lastly, the option *All Fields* for PubMed, were selected. When performing the search on all three databases, the following search string was applied (ACM result syntax output):
(+Chatbot* Training Education Apprenticeship Teaching)

In addition, the TeXMed [4] website was used to generate a BibTex file for PubMed. All extracted results were then managed by Zotero[4]. Besides an additional Excel file was used to summarize relevant information related to a set of predefined dimensions of interest. These dimensions comprise, for example, the technical solution such as the usage of the Artificial Intelligence Markup Language (AIML), Machine Learning (ML) algorithms and the implemented user interface (UI). Aspects related to an educational concept were also considered. The final update of all literature entries was conducted on 26th January 2018.

2.2. Technical Setup

All ML prototypes were developed either directly in Python (scikit-learn [5]) or with an adapter class for Rasa NLU [6]. Solutions that weren't available in Python were excluded. Python was used as it is the core language for many scientific fields including Artificial Intelligence (AI) & ML while still offering a highly readable code [7]. This allowed a more efficiently use of own software fragments and created a uniform approach for the overall evaluation using the inbuilt method *cross_val_score* [8] of scikit-learn for the LOO cross-validation. In addition, Rasa NLU was selected as an open source chatbot platform, which allows an execution on a private server. This can also be a future prerequisite, for example, if chatbots build up on sensitive data from patients or students. Therefore, external service providers such as Google or Facebook were excluded from evaluation. In order to use Rasa for classification, all anamnestic questions and the corresponding intents were specified in a separate file using the required markdown language. For this, each intent was described by a heading and the chatbot's knowledge base as an enumeration of all anamnestic questions in plain German. An example for a question and its intent is: "What medical complaints do you have?" (intent: *complaints_identification*). No entities and no sentence modifications are applied during this step for Rasa.

[2] https://ieeexplore.ieee.org/Xplore/home.jsp, last access: 29.01.2019.

[3] https://dl.acm.org/, last access: 29.01.2019.

[4] https://www.zotero.org/, last access: 29.01.2019.

Besides Rasa, different prototypes were also implemented in scikit-learn for classifying the intent. Thereby, each question was transformed into a high dimensional vector representation using the *Bag of Words (BOW)* approach [9] to generate the feature matrix (rows: anamnestic questions, columns: BOW generated words as features). In addition, methods like tf-idf, hashing and Doc2Vec were considered. However, BOW was selected as it is simple to use while showing sufficient accuracy for an initial proof of concept.

For the purpose of data exploration, additional visualizations were generated by Exploratory[5] and yellowbrick[6]. Finally, the evaluation process was conducted by the *cross_val_score* method of scikit-learn, applying it on all implemented prototypes. The specific test procedure for *cross_val_score* was set to LOO [8] and the individual scoring methods (micro-averaged) *recall*, *precision* and *f1* were selected [10]. Finally, all results were bundled into a Slack[7] application.

3. Results

3.1. Results of the Literature Review

The literature review led to 332 papers on IEEE, 124 papers on ACM and 2 papers on PubMed. They were then added to Zotero using the generated BibTex files, which also included abstracts if available. In summary, 458 papers were considered from these three databases using standard export and import of each database provider, TeXMed and Zotero. Nonetheless, several files were removed before or during the process of sighting each paper's title and abstract. The reason was either duplicates (5), no valuable information (24), e.g. referring only to a schedule [11] or having no access (1) [12]. Finally, 428 papers could be usefully sighted for title and abstract. At this stage, all chatbots were considered, which were integrated in a more advanced educational concept, e.g. in a concept of a Massive Open Online Course (MOOC). Therefore, results like answering FAQs of a university [13], supporting the degree program choice of a student [14] were not further reviewed. In addition, all results were excluded, which weren't enough focusing on a chatbot approach, e.g. only listing a chatbot as an example [15]. After sighting each title and abstract, 36 papers from IEEE, 7 from ACM and 1 from PubMed were sighted on their full text information, leading to 21 accepted publications (14 IEEE and 7 ACM papers). During this step, the paper quality itself wasn't considered as an additional criterion, only the chatbot context was decisive. In the following, a short summary of the results related to the educational setting is presented.

Frequently, chatbots were integrated in a MOOC scenario [16][17][18][19]. Thereby, Demetriadis et al. [16] focused on creating a more productive talk using transactive questions and conceptual links to shape the relevant domain model of a task. Kloos et al. [17] proposed a MOOC complementary chatbot, allowing to learn Java in several interaction modes, such as review and gaming. Besides MOOCs, conversational agents were also applied in Virtual Reality (VR) [20][21][22] creating an immersive educational environment. Tsaramirsis et al. [21], for example, simulated the experiences of a student in a classroom, including the communication with the lecturer. If the lecturer didn't

[5] https://exploratory.io/, last access: 29.01.2019.

[6] https://www.scikit-yb.org/en/latest/, last access: 20.03.2019.

[7] https://slack.com/, last access: 29.01.2019.

respond to a student's question, an inbuilt AIML Chatbot was used to generate the answer. Other chatbot realizations covered the use case of language learning [20][23][24]. Troussas et al. [23] developed a mobile chatbot for learning vocabularies through text or voice response. Further, gamification elements were incorporated by [25][26][27]. Pereira [25], for example, created a quiz chatbot for students in different subjects. Thereby, a Telegram UI was applied since students were familiar in using such instant messaging services [25]. Besides these results, Webber [28] was the only fitting VP approach that was referenced within the literature results. However, Webber builds on a rule-based SQL approach [29] that was published in 2005 [28]. Therefore, the intention of this paper is to revisit the concept of a VP chatbot, considering next to classical ML methods a modern approach (Rasa NLU) and a mobile chatting app (Slack) as it was suggested by Io and Lee [30] in a recent bibliometric analysis about chatbots.

3.2. Data Description and Exploration

Several data sources were integrated to train and evaluate the different prototypes. This included questions from the CAMPUS database and online resources [31][32][33][34]. They were used to generate data sets with the modified *textcorpus-generator*[8] project. Thereby, each anamnestic question was annotated by a single intent based on a personal assumption that derived from the gained insight of these resources. A typical question from the data set is for example "How much do you currently smoke per day?" (Intent: *Smoking*) or the one given in Section 2.2.

Table 1. Basic properties of the anamnestic questions corpora

	Features		
Dataset	**Number of words**	**Number of chars**	**Proportion of stop words**
AnamnesticData (n=109)	**(SD: 4.38)**	**(SD: 26.65)**	**(SD: 0.136)**
MED/AVG/MIN/MAX	8 / 9.3 / 3 / 21	46 / 52.52 / 16 / 135	28.6 / 29.3 / 0 / 57.14

The basic properties of the underlying questions corpora are described by several key features, being shown in Table 1. For each corresponding corpus, either the median (MED), average (AVG), minimum (MIN) or maximum (MAX) were calculated.

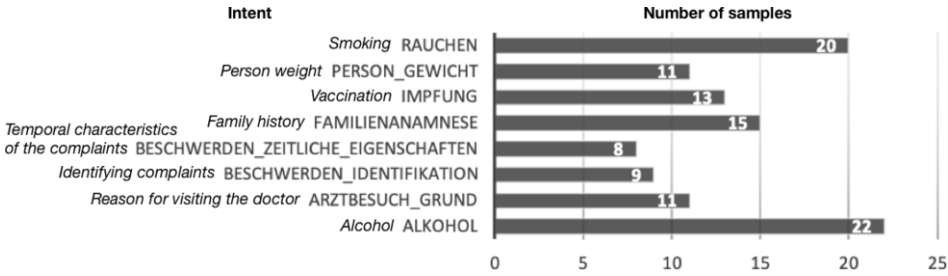

Figure 1. Distribution of all intents in given anamnestic questions.

Additionally, an absolute frequency distribution, described in Figure 1, gives an insight of the amount of annotated questions for each single intent. This data distribution

[8] https://github.com/pagesjaunes/textcorpus-generator, last access: 09.12.2018.

represents a general expected imbalance but doesn't claim to reflect the future reality. The underlying assumption is based on the opinion that certain intents won't allow to create the same amount of questions because they are either more specific (e.g. *Person weight*) or more generally designed (e.g. *Smoking*). A future study must therefore show how to define intents to avoid overlaps before ML classification.

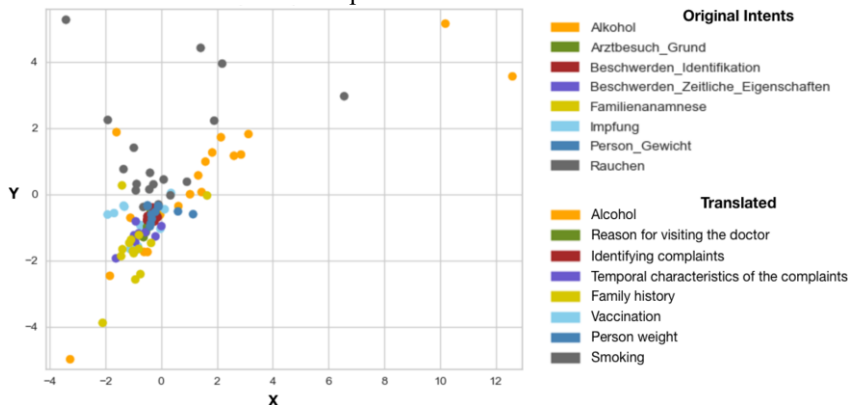

Figure 2. Embeddings of anamnestic questions in 2D vector space after PCA.

Figure 2 indicates this overlapping issue by using a 2D representation of each single anamnestic question. It was created by applying a Principal Component Analysis (PCA) on the BOW transformed questions. BOW generated a vocabulary with 385 words for all questions. For a future use, it is therefore crucial to find more intents like, for example, *Alcohol* and *Smoking* (Figure 2), which generate distinctive ML features (better distributed vectors). This would facilitate a clear ML classification and thus allowing to return the right VP answer for the medical student.

3.3. Implementing prototypes for intent classification

To determine the intent of an anamnestic question different prototypes were developed with scikit-learn and Rasa (Section 2.2). For scikit-learn, all anamnestic questions (n=109) were first transformed to a high-dimensional vector space using the BOW model (385 features) without applying any prior preprocessing. After sentence embedding various supervised ML algorithms were applied, which were already integrated in scikit-learn, including: *Random Forest* (RF), *Naïve Bayes* (NB), *Linear Support Vector Classification* (Linear SVC) and the *Logistic Regression* (LR). Thereby, the classification targets were set to the intents (n=8) while having the overall distribution described in Figure 1.

Figure 3. Slack integration of the anamnesis chatbot. User chooses Rasa NLU as a classifier. Input sentence contains an intentional spelling mistake and the chatbot returns the classified intent and its confidence value.

Finally, all created prototypes were bundled into a Slack application. Figure 3 illustrates the final Slack UI using the Rasa prototype for intent classification. Thereby, each single prototype is selectable by using the keyword: *set_mode_x* where *x* stands for the number of the individual prototype, e.g. *x=2* for Rasa (Figure 3). All additional mappings can be listed by using the chatbot's *help* command. All in all, the use of Slack creates a UI, which allows the communication between student and VP through a modern, well-known messaging service.

3.4. Classifier Evaluation

Table 2. Performance measures for the implemented approaches.

	Scoring method		
Approach	**recall_micro in %**	**precision_micro in %**	**f1_micro in %**
Platform			
Rasa*	88.073 ± 32.410	88.073 ± 32.410	88.991 ± 31.300
Scikit-Learn			
RF*	85.321 ± 35.390	84.404 ± 36.282	82.569 ± 37.938
NB	81.651 ± 38.706	81.651 ± 38.706	81.651 ± 38.706
LSCV	88.991 ± 31.300	88.991 ± 31.300	88.991 ± 31.300
LR	85.321 ± 35.390	85.321 ± 35.390	85.321 ± 35.390

***** score value fluctuates at over 80%

Since the total number of the available data for a specific intent was small, the LOO cross-validation method was selected as a test procedure. In the next step each scikit-learn prototype inherited the *BaseEstimator*[9] class and implemented the adapter methods *fit(self, X, y)* and *predict(self, X)*. Subsequently, an object of this class was passed to the *cross_val_score* method to perform the final evaluation. Thereby, the following scoring methods (micro-averaged) were applied: *recall*, *precision* and *f1*. The results of this respective approach are listed in Table 2, where each value represents an average of all *n* measurements. All prototypes achieve a score of over 80% for each individual combination with Rasa and LSVC delivering the best results (RF excluded because of high fluctuation). However, since the SD of each score is very high due to LOO, all current prototypes (Section 3.2) should remain selectable by the UI (Section 3.3) for a future field study. This would allow to record not only new real data for cross-validation but also gain feedback on the individual perception of use by each medical student. Both insights could then be analyzed to select the final prototype for the use in a case-based learning system.

4. Discussion

Considering the given data that is shown in Table 1, both mean values indicate that the total number of words and characters in the training data is low. Instead, the proportion of stop words with about 29% is quite high. Bearing all this in mind, the subsequent ML

[9] https://scikit-learn.org/stable/modules/generated/sklearn.base.BaseEstimator.html, last access: 12.02.2019.

algorithms had only few valuable information to deal with and still performed fine (>80%, but high SD due LOO). Further improvements could be made by ML parameter optimizations or by increasing the data quality when allowing medical students or experts to ask questions and subsequently integrate their feedback for training purposes.

The results of Table 2 also show that there are only minor differences in terms of classification performance between the Rasa NLU platform and the assembled scikit-learn implementations. It would be interesting to investigate whether lemmatizing or additional feature extractions, e.g. from grammatical structures, could lead to further performance improvements for the scikit-learn prototypes.

The current version of the Slack UI could allow an unstructured anamnesis survey between chatbot and user by replacing the intent with a VP answer. Thereby, it can be used on the computer as well as within a mobile application. For future development, the dialog system could be extended by storylines, e.g. by Rasa stories, making the overall conversation more sophisticated. At this point, additional concepts like voice commands [20] or a VR avatar [21] can be also considered. If there are no further improvements in data quality the interaction between user and chatbot might be facilitated by additional UI elements like in Fadhil and Villafiorita [26]. This could help building a feedback channel, e.g. displaying a small set of intents for user selection if the confidence of the chatbot isn't sufficiently high enough. As a result, the user could indicate a suitable intention or deny the given suggestion completely. These statements could then be forwarded to an author's Slack workspace for revision and re-added to the chatbot for Reinforcement Learning.

References

[1] M. Riemer and M. Abendroth, Virtuelle Patienten: Wie werden sie aus Sicht von Medizinstudierenden am besten eingesetzt?, *Ger. Med. Sci. GMS E-J.*, (2013).

[2] M. R. Fischer et al., Virtuelle Patienten in der medizinischen Ausbildung: Vergleich verschiedener Strategien zur curricularen Integration, Z. Für Evidenz Fortbild. Qual. Im Gesundheitswesen, **102**(10), (2008), 648–653.

[3] PRISMA, PRISMA Flow Diagram, http://prisma-statement.org/prismastatement/flowdiagram.aspx, last access: 29.01.2019.

[4] TeXMed, TexMed – a BibTeX interface for PubMed, https://www.bioinformatics.org/texmed/, last access: 29.01.2019.

[5] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, **12**, 2011, 2825–2830.

[6] Rasa NLU, Rasa NLU: Language Understanding for chatbots and AI assistants, https://rasa.com/docs/nlu/, last access: 30.01.2019.

[7] G. Rashed and R. Ahsan, Python in Computational Science: Applications and Possibilities, *Int. J. Comput. Appl.*, **46**(20), 2012, 26-30.

[8] Scikit-learn, API Reference, https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection, last access: 29.01.2019.

[9] Scikit-learn, Text feature extraction, https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction, last access: 22.01.2019.

[10] Scikit-learn, Model evaluation: quantifying the quality of predictions, https://scikit-learn.org/stable/modules/model_evaluation.html, last access: 29.01.2019.

[11] IEEE Xplore, Schedule, *2018 Zooming Innovation in Consumer Technologies Conference (ZINC)*, (2018).

[12] S. Garg et al., Clinical Integration of Digital Solutions in Health Care: An Overview of the Current Landscape of Digital Technologies in Cancer Care, *JCO Clin Cancer Inf.*, **2**(2), 2018, 1-9.

[13] B. R. Ranoliya et al., Chatbot for university related FAQs, *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, 1525–1530.

[14] S. Mirri et al., User-driven and open innovation as app design tools for high school students, *in 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC),* 2018, 6–10.
[15] C. Kuo and J. Z. Shyu, An Innovative Syndicate Medium Ecosystem, *in 2018 IEEE International Symposium on Innovation and Entrepreneurship (TEMS-ISIE)*, 2018, 1–5.
[16] S. Demetriadis u. a., „Conversational Agents as Group-Teacher Interaction Mediators in MOOCs“, in 2018 Learning With MOOCS (LWMOOCS), 2018, S. 43–46.
[17] C. D. Kloos et al., Design of a Conversational Agent as an Educational Tool, *in 2018 Learning With MOOCS (LWMOOCS),* 2018, 27–30.
[18] H. Hsu and N. Huang, Xiao-Shih: The Educational Intelligent Question Answering Bot on Chinese-Based MOOCs, *in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, 1316–1321.
[19] A. Mitral et al., MOOC-O-Bot: Using Cognitive Technologies to Extend Knowledge Support in MOOCs, *in 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2018, 69–76.
[20] A. Berns et al., Exploring the Potential of a 360° Video Application for Foreign Language Learning, *in Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, New York, 2018, 776–780.
[21] G. Tsaramirsis et al., Towards simulation of the classroom learning experience: Virtual reality approach, *in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, 1343–1346.
[22] I. Stanica et al., VR Job Interview Simulator: Where Virtual Reality Meets Artificial Intelligence for Education, *in 2018 Zooming Innovation in Consumer Technologies Conference (ZINC),* 2018, 9–12.
[23] C. Troussas et al., Integrating an Adjusted Conversational Agent into a Mobile-Assisted Language Learning Application, *in 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2017, 1153–1157.
[24] X. L. Pham et al., Chatbot As an Intelligent Personal Assistant for Mobile Language Learning, *in Proceedings of the 2018 2Nd International Conference on Education and E-Learning*, New York, 2018, 16–21.
[25] J. Pereira, Leveraging Chatbots to Improve Self-guided Learning Through Conversational Quizzes, *in Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, New York, 2016, 911–918.
[26] A. Fadhil and A. Villafiorita, An Adaptive Learning with Gamification & Conversational UIs: The Rise of CiboPoliBot, *in Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, New York, 2017, 408–412.
[27] K. Katchapakirin and C. Anutariya, An Architectural Design of ScratchThAI: A Conversational Agent for Computational Thinking Development Using Scratch, *in Proceedings of the 10th International Conference on Advances in Information Technology*, New York, 2018, 7:1–7:7.
[28] G. M. Webber, Data Representation and Algorithms For Biomedical Informatics Applications, PhD thesis, Harvard University, 2005.
[29] A. S. Lokman, J. M. Zain, F. S. Komputer and K. Perisian, Designing a Chatbot for diabetic patients*, International Conference on Software Engineering & Computer Systems*, 2009.
[30] H. N. Io and C. B. Lee, Chatbots and Conversational agents: A bibliometric analysis, in *IEEE International Conference on Industrial Engineering and Engineering Management*, 2017, 215-219.
[31] Jairvargas, Anamnese, https://www.slideshare.net/jairvargas/anamnese-44468748, last access: 29.01.2019.
[32] Alk-info.com, Alkoholtest mit 22 Fragen, Schnell-Test auf Alkoholgefährdung, https://www.alk-info.com/tests/print/439-alkoholtest-mit-22-fragen-schnell-test-auf-%20alkoholgefaehrdung,last access: 21.03.2019.
[33] U. Latza et al., Erhebung, Quantifizierung und Analyse der Rauchexposition in epidemiologischen Studien, Robert Koch Institut, 2005.
[34] Robert Koch Institut, Journal of Health Monitoring – Fragebogen zur Studie „Gesundheit in Deutschland aktuell”(GEDA2014/2015-EHIS), https://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GBEDownloadsJ/Supplement/JoHM_2017_01_gesundheitliche_lage9.pdf?__blob=publicationFile ,last access: 11.02.2019.