Hong Li, Tamar Arndt and Miloš Kravčík

Improving Chatbots in Higher Education

Intent Recognition Evaluation

Intelligent assistants provide a lot of potential in the development of scalable mentoring solutions for higher education. Chatbots became popular also in this field, but they may cause users' frustration if they do not understand the user correctly. Therefore, it is crucial to evaluate their performance in recognizing user messages, in order to know their weak points and improve them accordingly. In our experiments, we took one chatbot used by students of educational sciences and evaluated it from various perspectives. The results indicate that this kind of validation can help to improve the usability of chatbots in the learning domain.

1. Introduction

In higher education (HE), learning is an individual process of actively acquiring and constructing knowledge, accompanied by mentoring by teachers. Feedback is a key impact factor for learning success if it is immediate and precise. In addition to cognitive factors, metacognitive, emotional and motivational aspects also play a crucial role in learning and mentoring. Compared to coaching and tutoring, the mentoring process is more spontaneous and holistic, based on the needs and interests of the mentee and focuses on psychological support. The mentoring relationship is more complex, interactive and based on emotions.

Digital learning environments offer potential to provide students with a wide range of support. Since higher education institutions work with limited resources, socio-technical infrastructures must be carefully designed in order to be able to scale supporting processes with the help of distributed artificial intelligence (Klamma et al., 2020). The available information technology can analyse the extensive learning data sets from the system logs, sensors and texts in order to reveal various aspects of learning progress and, if necessary, the need for intervention. The aim is to relieve the teachers and at the same time to maintain the quality of the teaching. It is important that the learners are in control and decide for themselves which data are to be made available for which purposes.

Intelligent conversation assistants, such as chatbots, can be used as a key component in a digital learning environment. A chatbot is a software application enabling an online chat conversation with a human. Intelligent chatbots make use of all kinds of artificial intelligence (AI), including natural language understanding (NLU), machine learning and deep learning. Chatbots are already successfully used in various different domains (e.g. as customer service). Also in the context of mentoring, the conversational nature and other characteristics of chatbots can add value. For example, they offer opportunities to create individual learning experiences (Winkler & Söllner, 2018). The ability to interact via natural language makes them intuitive to use. In this context, an essential quality requirement of chatbots emerges: they must be able to interpret users' messages and intents correctly in order to avoid frustration. To that end, in this paper, we describe our approach to analyse and improve a chatbots' model for NLU.

In the following, we briefly introduce the related work. Then we describe the methodology used. The main part presents our case study with several evaluation experiments. At the end we summarize our contributions.

2. Related Work

Intelligent assistants can be used in HE both for administrative and learning support, for example for automatically replying to students on behalf of the academic staff (Hien et al., 2018). Chatbots can enhance students' learning experience and facilitate the achievement of student-centred learning. They can conduct research, mark online exams and assist students to communicate well (Sandu & Gide, 2019).

HE chatbots use AI technology for supporting learning at scale by automatically answering a variety of routine, frequently asked questions, and automatically replying to student introductions. Their design can gradually move from using an episodic memory of previous question-answer pairs to using semantic processing based on conceptual representations (Goel & Polepeddi, 2016). To avoid frustration of users, it is reasonable to use a certain confidence threshold for automatic interventions.

Chatbots seem to be used mainly as answering-machines, but not for assignment individualization purposes so far (Bollweg et al., 2018). Nevertheless, automated assessment can be facilitated by a chatbot that can grade students' responses to generated questions on a similar level as a human instructor (Ndukwe et al., 2019).

In the tech4comp project[1] we aim at personalized competence development through scalable mentoring processes, which is investigated at several German universities in three different domains – educational sciences, mathematics and informatics. An important limitation of mentoring in higher education are the available resources, but the benefits of individual feedback can also be achieved through the use of appropriate technology. As an interface for both mentors and mentees, we use intelligent mentoring chatbots, tailored for mentoring processes and integrated with external learning applications (Neumann et al., 2021). The implementation of such chatbots is based on Natural Language Understanding (NLU) to identify the respective intent of the student and to respond properly.

In our research, we were looking for studies dealing with intent recognition of chatbots in HE, but we have not found such specific evaluations. Nevertheless, we consider such validation important and want to share our experience, helping to bridge the gap between developers and practitioners.

[1] https://tech4comp.de

3. Methodology

In this paper, we describe our approach to analysing and improving the intent recognition and NLU of a chatbot to support iterative evaluation and development of chatbots. We do this using a chatbot as an example, which has been developed in the tech4comp project to support students in teacher training (Neumann et al., 2021). To build the chatbot, we use the Social Bot Framework (Neumann et al., 2019) that integrates Rasa[2] for the intent recognition, which is an AI platform for personalized conversations at scale. Rasa includes an open-source NLU framework and together with a toolset for improvements of virtual assistants it supports the creation of chatbots. The Rasa NLU component comprises coupled modules combining a number of natural language processing and machine learning libraries. It trains a classifier model using a list of annotated statements with intents as training data. With the trained model, for each input message, the Rasa NLU server can give a list of intent and confidence pairs, and the intent with highest confidence is chosen as recognition result (Bocklish et al., 2017). Definition of intents is a crucial part of a Rasa chatbot, which includes the phrases that are expected from the user. One *intent* can be expressed by various *examples*.

Our chatbot evaluation consists of three parts, which we describe next:

- Validation of intent recognition,
- Analysis of conversations,
- Improvement of the Rasa-NLU-Model.

3.1 Validation of Intent Recognition

It is helpful to validate the definition of intents, in order to improve the chatbot's performance. This can be done before users start to use the chatbot. Rasa NLU classifies the user messages into user intents, which is done with a certain *confidence*. In this way (semantically) similar intents can be found. It is worth analysing when this parameter reaches low values, identifying problematic intents. This includes the identification of the top two intents in the intents-list for each sentence example and an analysis of the confidence difference.

In this context, another important parameter is the *precision*, which helps to identify errors. It is the ability of the classifier not to label a sample that is negative as positive. The precision is the ratio *tp / (tp + fp)* where *tp* is the number of true positives and *fp* the number of false positives. The best value is 100 % and the worst value is 0 %. When an intent has a low precision, we need to identify what is wrong. For instance, there can be duplicates of some examples in different intent definitions.

[2] https://rasa.com

3.2 Analysis of Conversations

Another approach takes into account the data from the chatbot logs and the corresponding analysis of conversations with real users. Here, we can find the distribution of the recognized intents, the distribution of the confidence and the statistics of the conversation path. If we add manual annotations by experts, another level of the analysis can be achieved. For such annotations, we considered sentences with a low intent recognition confidence and also randomly selected sentences for each intent. The experts approved the correct recognitions and in case of incorrect ones they also suggested the right solutions. Such annotations were used in our evaluation. Regarding the annotation, being an expert means to be familiar with the content and functional scope of the chatbot and thus being able to interpret the real user's intention from a message and to check it against the intent recognized by the NLU-Model.

3.3 Improvement of the Rasa-NLU-Model

We can improve the Rasa-NLU-Model if we use the annotations made by experts and enhance the intent examples with suitable annotated sentences. To investigate the effectiveness of this improvement, we evaluate the new trained model regarding the distribution of the recognized intents as well as the distribution of their confidence.

4. Case Study

To describe our approach, we use a chatbot offered in a seminar attended by about 800 students in teacher training each semester. Self-study reading of the seminar literature plays a major role there. To support self-study, the chatbot gives writing tasks on the literature and provides feedback on the written texts sent by students. The feedback is generated using T-MITOCAR, a computational linguistic text analysis software that analyses the text structure and generates its graph visualization (Pirnay-Dummer and Ifenthaler, 2011). The automatically generated feedback is sent back to the student after submission along with a brief explanation. The main functions of the chatbot are showing current writing tasks, providing information on how to submit a text, accepting submissions, and sending feedback. These are covered by the intents *show-tasks* and *submission*. The chatbot can answer general questions and make a little small talk, for which it understands the intents *greet* and *goodbye*. The intent *badbehavior* was added to react to offensive language in the chat. When the chatbot recognizes the intent *credit*, it gives information on how a writing task is being assessed. The intent *tmitocar* refers to background information on how the feedback is generated. The intent *privacyanddata* informs how personal data is handled by the chatbot. For the intent *functions,* the chatbot gives its functionality overview. The intent *contact* is there to point to further contact options, such as a support forum. The first version of the chatbot was provided in the winter semester 2020/2021 and a second version in the summer semester 2021. We validated both versions of the chatbot in several experiments, using the Rasa NLU component.

4.1 Validation of Intent Recognition

4.1.1 Experiment 1.A

In this experiment, the chatbot training data from the winter semester 2020/2021 were validated, in order to find out whether the intents were defined properly. It included 14 intents and 203 example sentences. This was used as a training set for the validation of intent recognition in the Rasa NLU. The classifier correctly labelled 200 out of 203 sentences, which makes the precision of 98,52 %. Fig. 1 shows the validation results and indicates two intents with a potential for improvement: 3 incorrectly labelled sentences were associated with 2 low confidence intents, where 3 duplicates were found. In these cases, the confidence did not exceed 0,5.

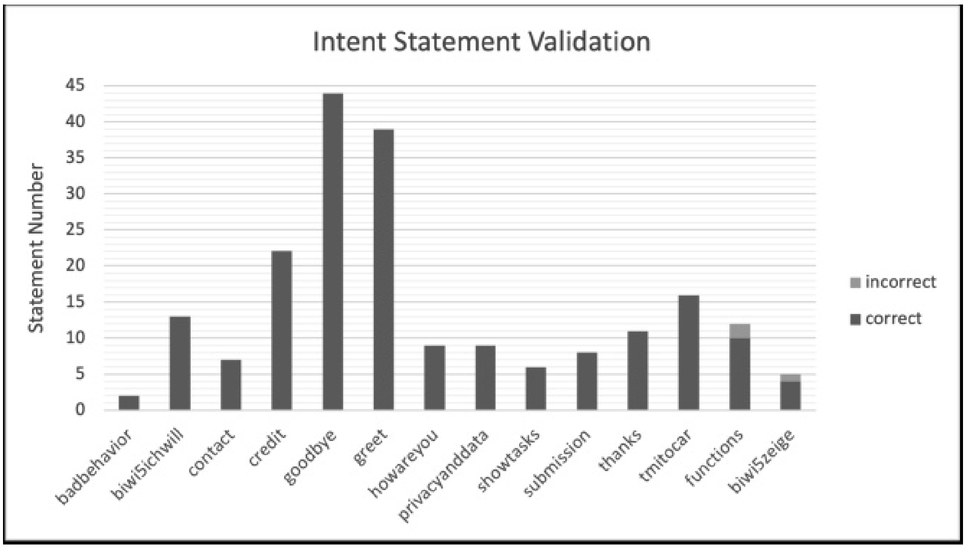

Figure 1: The validation results of the training data

4.1.2 Experiment 1.B

This experiment took place in the summer semester 2021, when 19 intents and 487 example sentences (max. 88, min. 7 per intent) were included. This training set was evaluated by the Rasa NLU. The precision reached 100 %. The lowest confidence was 0,375, which was found in one sentence (intent: *contact*, text: "Du hast meine Frage nicht beantwortet"). Also in other 52 cases the confidence was below 0,5. The semantic similarity of some intents was caused by a lack of examples.

4.2 Analysis of Conversations

To illustrate conversations between the student and the chatbot, here is a short example:

- *Student:* ”Bitte gib mir eine Schreibaufgabe” *[intent: showtasks]*
- *Chatbot:* ”Aktuell kannst du die Schreibaufgaben zu *Themenblock I (Aufgaben 1–3)* und *Themenblock II (Aufgaben 4–6)* bei mir einreichen. Abgaben zu Themenblock I kannst du bis 06.12. bei mir hochladen. (…) *Ziehe zum Hochladen dein Dokument in dieses Chatfenster.*”

4.2.1 Experiment 2.A

This evaluation took place in the winter semester 2020/2021, more precisely between October 26 and November 30 (36 days). 575 conversations with a unique user were considered, with the average length of 24 messages (maximum 110, minimum 1). The number of messages was 13.824 (9.144 from the bot, 4.680 from the students). These include 2.579 messages from students with no text, as these were uploads of documents. However, students wrote 72 multi sentence messages. The Rasa intent recognition module analysed 2.015 text messages, from which 1.396 were distinct ones. Fig. 2 shows the distributions of intent recognition in conversations (without document upload).

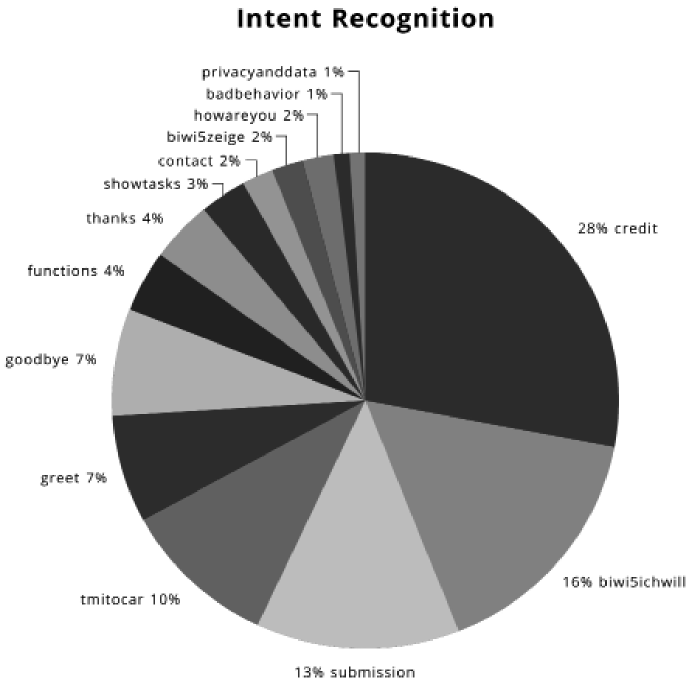

Figure 2: Intent distribution in conversations

For the chat path analysis, we identified 4.447 user-bot message pairs and mostly (99,28 %) the bot responded in 1 s, with the average response time 144 ms. Conversations between the bot and the (unique) user were cut into threads (sessions) if they included breaks of more than 5 min. In 490 conversations without noise we identified 2.293 threads, with the average 4,7 threads per conversation (maximum 16, minimum 1). This corresponds with the task to submit 5 written texts in the considered time period. On average, there were 4,8 messages per thread (maximum 30, minimum 1). Users mostly started and ended a session with a document upload. In 2.293 threads 327 different paths (158 with one step) have been identified and the highest frequency (55 %) had a simple document upload. Fig. 3 visualizes the one step path analysis, distinguishing various categories of frequencies. The central action is apparently the document upload, which is often done repeatedly, without other interventions.

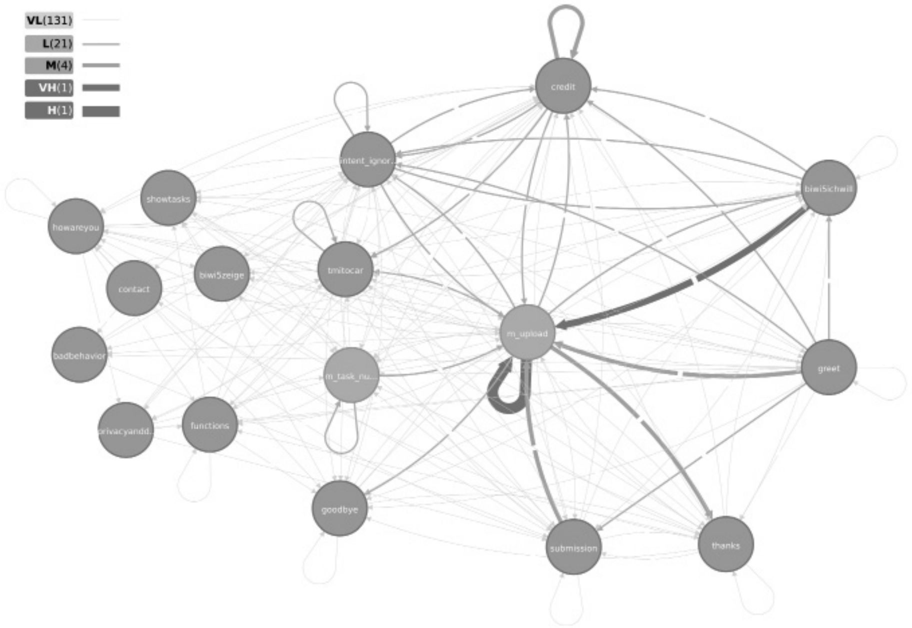

Figure 3: Chat path analysis (VH>200, H>100, M>50, L>10, VL others)

4.2.2 Experiment 2.B

For the manual annotation in October – November 2020, we selected 845 sentences – a part of them with the lowest confidence and the rest randomly, considering an even distribution per intent. The annotator evaluated the Rasa NLU intent recognition results. If the recognition was incorrect, the annotator could give correct suggestions. The Gold-Standard Corpus was built from two parts: 1. sentences annotated as correct, 2. sentences annotated as incorrect, with suitable suggestions. This evaluation led to the Gold-Standard Corpus with 712 examples for 21 intents (12 old, 8 new, 1 default). The intents with the highest number of examples were: *default* 179, *submission* 125 and *showtasks* 88 (Fig. 4).

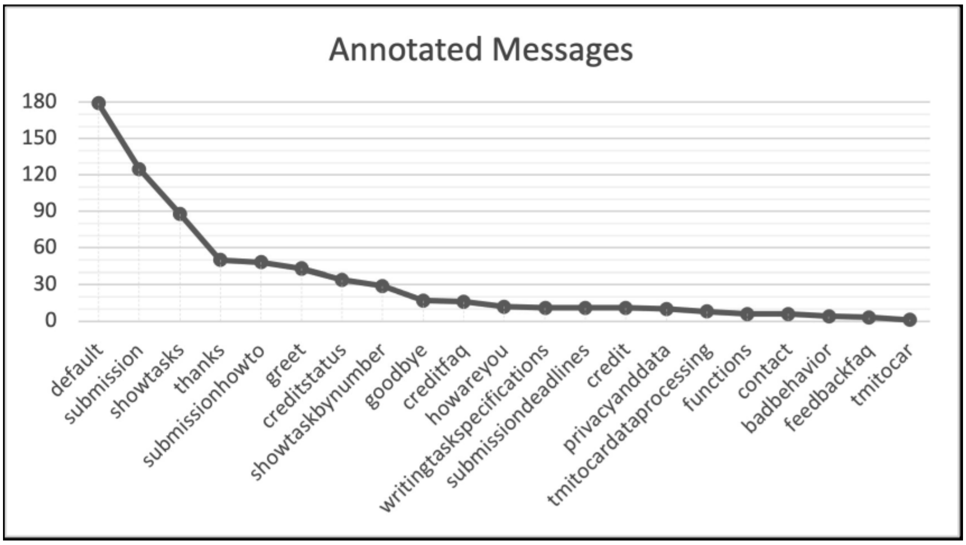

Figure 4: Number of examples per intent in the Gold-Standard Corpus

4.2.3 Experiment 2.C

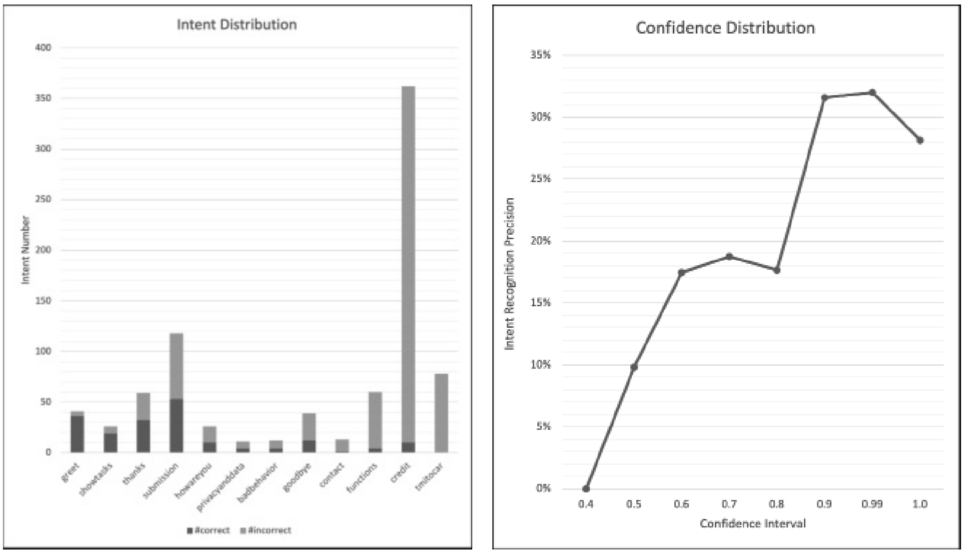

Figure 5: Intent and confidence distribution of conversations

The annotations were used to evaluate the intent recognition results. From 845 annotated messages, only 185 were correct (precision 21,89 %). Fig. 5 shows the corresponding intent and confidence distributions. Among the 12 old ones, the intents *greet* and *showtasks* have the best precision (88 % and 73 %) – perhaps because they have most example statements. The precision of the recognition result with higher confidence values is also much better than the one with low confidence values: the confidence value can be used as an indicator to evaluate the performance of the intent recognition.

4.3 Improvement of the Rasa-NLU-Model

4.3.1 Experiment 3

The acquired log data can be used as a training corpus to improve the Rasa-NLU-Model of the chatbot. We used the logs from the previous experiment collected between December 1st, 2020 and February 10th, 2021. In this data 681 conversations of unique users with the bot were identified, with an average of 41 messages (max. 127, min. 4). In total 28.218 messages were found. The new intent training set was created from the previous one and enhanced with the annotated Gold-Standard from Experiment 2. The number of intents remained 12, but the number of examples increased from 183 to 550. With each of the training sets, more than 900 distinct text messages from students have been used. Fig. 6 compares the two Rasa NLU models, showing the number of messages recognized with the corresponding confidence. The

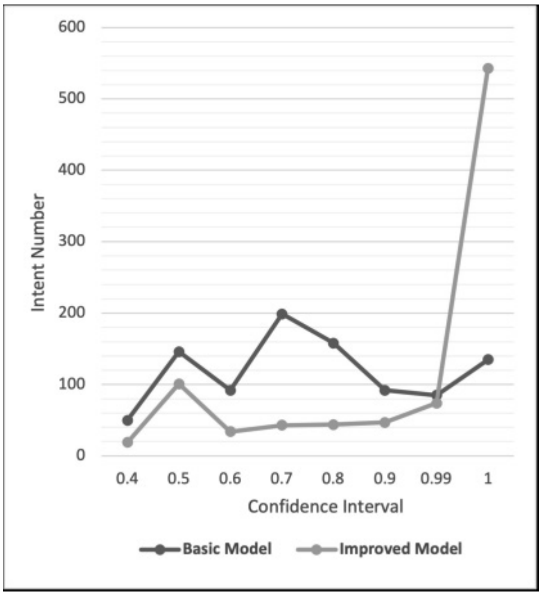

Figure 6: Confidence distribution of the intent based on two models

improvement is quite apparent: 543 messages (60 %) recognized with the highest confidence (above 0.99).

5. Conclusion

Chatbots have the potential to support a whole range of tasks in HE, especially when integrated with other tools. But to bring reasonable benefits, they have to achieve a sufficient quality, which needs to be evaluated. In this paper, we presented several ways how a chatbot developed in the Rasa platform can be tested and improved. The validation of intent recognition revealed a couple of intents that could be improved in the basic set. In the extended version more semantically similar intents were identified, suggesting a definition of additional examples. The next analysis showed the distribution of recognized intents in real conversations as well as the typical chat paths used by students. Manual annotation of selected sentences helped to create a new corpus and perform further evaluations. These annotations together with the log data helped to improve the confidence in our Rasa-NLU-Model quite dramatically. The results show that these approaches can lead to continuous improvements of provided services. Further developments can take into account more contextual information as well as sentiment analysis, as motivation and emotions are crucial for mentoring in HE.

Acknowledgements

The authors would like to thank the BMBF for their kind support within the project "Personalisierte Kompetenzentwicklung durch skalierbare Mentoringprozesse" (tech4comp) under the project id 16DHB2110.

Literature

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Bollweg, L., Kurzke, M., Shahriar, K. A., & Weber, P. (2018). When robots talk – improving the scalability of practical assignments in MOOCs using chatbots. *EdMedia+ Innovate Learning* (pp. 1455–1464). AACE.

Goel, A. K., & Polepeddi, L. (2018). Jill Watson: A virtual teaching assistant for online education. In *Learning Engineering for Online Education* (pp. 120–143). Routledge. https://doi.org/10.4324/9781351186193-7

Hien, H. T., Cuong, P. N., Nam, L. N. H., Nhung, H. L. T. K., & Thang, L. D. (2018). Intelligent assistants in higher-education environments: the FIT-EBot, a chatbot for administrative and learning support. *Proceedings of the ninth international symposium on information and communication technology* (pp. 69–76). https://doi.org/10.1145/3287921.3287937

Klamma, R., de Lange, P., Neumann, A. T., Hensen, B., Kravcik, M., Wang, X., & Kuzilek, J. (2020). Scaling Mentoring Support with Distributed Artificial Intelligence. *Interna-*

tional Conference on Intelligent Tutoring Systems (pp. 38–44). Springer, Cham. https://doi.org/10.1007/978-3-030-49663-0_6

Ndukwe, I. G., Daniel, B. K., & Amadi, C. E. (2019). A machine learning grading system using chatbots. *International Conference on Artificial Intelligence in Education* (pp. 365–368). Springer, Cham. https://doi.org/10.1007/978-3-030-23207-8_67

Neumann, A. T., de Lange, P., and Klamma, R. (2019). Collaborative Creation and Training of Social Bots in Learning Communities. *IEEE 5th International Conference on Collaboration and Internet Computing (CIC)* (IEEE), (pp. 11–19). https://doi.org/10.1109/CIC48465.2019.00011

Neumann, A. T., Arndt, T., Köbis, L., Meissner, R., Martin, A., de Lange, P., Pengel, N., Klamma, R. & Wollersheim, H. W. (2021). Chatbots as a Tool to Scale Mentoring Processes: Individually Supporting Self-Study in Higher Education. *Frontiers in Artificial Intelligence*, *4*, 64. https://doi.org/10.3389/frai.2021.668220

Pirnay-Dummer, P. and Ifenthaler, D. (2011). Reading guided by automated graphical representations: How model-based text visualizations facilitate learning in reading comprehension tasks. *Instructional Science* 39, (pp. 901–919). https://doi.org/10.1007/s11251-010-9153-2

Sandu, N., & Gide, E. (2019). Adoption of AI-Chatbots to enhance student learning experience in higher education in India. *18th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1–5). IEEE. https://doi.org/10.1109/ITHET46829.2019.8937382

Winkler, R. & Söllner, M. (2018). Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. Academy of Management Annual Meeting (AOM). https://doi.org/10.5465/AMBPP.2018.15903abstract