Intent Identification and Analysis for User-centered Chatbot Design: A Case Study on the Example of Recruiting Chatbots in Germany

Stephan Böhm,
Judith Eißer,
and Sebastian Meurer

CAEBUS Center of Advanced E-Business Studies
RheinMain University of Applied Sciences
Wiesbaden, Germany
e-mail: {stephan.boehm, judith.eisser,
sebastian.meurer}@hs-rm.de

Olena Linnyk, Jens Kohl,
Harald Locke, Levitan Novakovskij,
and Ingolf Teetz

Milch & Zucker AG,
Gießen, Germany
e-mail: {olena.linnyk, jens.kohl,
harald.locke, levitan.novakovskij,
ingolf.teetz}@milchundzucker.de

***Abstract*—Chatbots are text-based dialogue systems that automate communication processes. Instead of communicating with a person, the user communicates with a computer system. Due to the use of Artificial Intelligence (AI) methods, such systems have become increasingly powerful in recent years and allow for more realistic dialogue processes. In particular, methods from the field of machine learning have contributed to an improved understanding of natural language. Nevertheless, such systems are not yet able to acquire the knowledge required to answer user queries independently. Dialogue structures and elements need to be defined as the conversational design of the chatbot. Herein, an user intent describes an information need or a goal that the user aims to achieve by entering text. For a user-centered chatbot design, a relevant set of intents must be identified and structured. In addition, training questions are required in order train the AI models for matching user input with the defined set of user intents. This article describes the procedure for developing chatbots using the example of an application in recruiting. The focus is on the appropriate identification and analysis of user intents. In our case study, the procedure for user-centered intent identification is described as well as approaches for the analysis and consolidation of intents. Furthermore, it is shown how corresponding measures affect the quality of intention identification.**

Keywords–Chatbots; Conversational Design; Prototyping; User Intent Analysis; User-centered Design; Machine Learning.

I. Introduction

The mode of communication has changed. Where in the past, information was normally only provided by companies in one direction and in a unidirectional one-to-many approach, interactive one-to-one dialogues are possible at large scales today [1]. Stakeholders can converse with companies and vice versa. Chatbots are a way to automate this dialogue process and are implemented to address this need [2]. Based on pattern matching and natural language processing methods or artificial intelligence, chatbots are automated dialogue systems for conversational scenarios [3]. They are utilized to mimic unstructured natural language dialogues normally prevailing in human-human conversations; either based on hand-built rules or on corpus-based AI functionalities, where data is mined from existent human-human conversations [4]. The potential of chatbots is vast and its diffusion continues to progress: According to a global chatbot market report by Research and Markets, the chatbot market size will be worth 9.4 billion US dollars by 2024 at an estimated compound annual growth rate of almost 30 percent [5]. Established in the 1960s, technological advancements constantly improved the technology so that today, chatbots hold the potential to support various business processes [6]–[8]. Especially in repetitive scenarios like answering Frequently Asked Questions (FAQ), AI-based technology, such as chatbots, are implemented to increase efficiency by improving quality while reducing costs [9].

This paper is about the implementation of chatbot solutions in recruiting, a special field of human resources that deals with finding and hiring new personnel for employers like companies and other institutions. In recruiting, chatbots can be deployed to transfer information to potential candidates and talents before, throughout and after the application process. They can be utilized to answer general questions regarding a certain position or the application process for example [6]. Through automation and the deployment of artificial intelligence functionalities, the processes of applicant sourcing and screening can be supported and the aspect of human bias in recruiting can be reduced [9]. In the current ”war for talents”, state-of-the-art technology enabling or at least facilitating the process of recruiting the most suitable talents at the most suitable points of contact for them is essential for organizational success and the formation of a competitive advantage [10]. There are several relevant and interesting use cases along the recruiting process, which can be supported by chatbot functionalities; the focus areas of this study will be shed light on when regarding recruiting chatbots in more detail in Section II-C.

The use of chatbots in recruiting is still relatively new. There are already many example applications (e.g., [11]), but these are often early pilot and test applications and in many cases not yet in permanent productive use. Nevertheless, there are more and more developers of chatbot solutions [12] and many of them use AI to promote such new applications. For decision makers in the HR sector with less technical experience, the impression sometimes arises that chatbot solutions are largely autonomous learning systems that only need to be

ISBN: 978-1-61208-829-7

implemented in companies and then acquire the knowledge to answer user questions themselves. However, this is a major misunderstanding. The use of AI in many chatbot frameworks is still largely limited to Natural Language Understanding (NLU) and the classification of user questions to predefined user intentions. Usually, however, the user intentions have to be created in the system and linked to certain actions for output. Developers of chatbots must therefore not only implement such solutions technically, but also define and structure dialogue contents in a conversational design [13]. The selection of the user intentions to be considered plays a special role, as it defines the application domain within which a chatbot can answer user requests in a meaningful way. For the identification and further analysis of such user intentions, however, the literature contains hardly any practical descriptions of the procedure [14].

This study regards the necessity as well as the actual formation process of a suitable intent set for a corpus-based recruiting FAQ chatbot while challenging the newly trained version against the former version of the dialogue technology prototype. After this introduction, an overview of the theoretical background is given in Section 2. Related work and studies are discussed before defining the research objectives of the study at hand in Section 3. The study's outline is presented in Section 4 with the methodology and the case study approach. Section 5 deals with the case study findings and its theoretical as well as practical implications. The last Section 6 presents final conclusions, limitations and an outlook on further research.

II. Research Background

In order to understand the problem of identification and analysis of user content, a brief discussion of some background information on conversational design will be given in the following. Afterwards, the technical implementation of AI-based chatbot solutions and the importance of training data will be discussed. The section concludes with an introductory description of FAQ chatbots in general and their application in connection with applicant tracking systems used in recruiting.

A. Conversational Design

Chatbots belong to the conversational interfaces [15]. Conversational interfaces are a special kind of interactive user interface, which enables a dialogue in natural language between humans and computers and can process user input as text or speech, oftentimes based on AI functionalities [13][16]. Popular conversational interfaces are voice assistants that react to spoken user input and chatbots, which are discussed here. In chatbot solutions, conversation typically takes place through typed text input and a front-end that can be, for example, embedded in a website or messaging solution [17]. Conversational design as a special discipline of interactive design deals with all tasks of designing conversational interfaces (e.g., stakeholder and goal definition, conversational flow design [16], actual development and testing) with the goal to provide a good user experience [18].

Like other objects of interactive design, chatbots have different design elements. The design of the front-end user interface is less in focus, since text input leaves little room for variation. Fist of all, interfaces for text input can be varied by the colours or by font characteristics. Furthermore, decisions on the chatbot's personality in the form of a specific persona [15][18] (e.g., use of avatars) or the use of graphic elements for the chatbot output such as buttons, images (moving or static) as well as emoticons and emojis can be considered as design aspects [19]. The tonality of the language is another exemplary design aspect [6]. The core of the chatbot's conversational design, however, is more concerned with determining the dialogue content and its logical structure. However, the respective design options for chatbot development are determined by the particular chatbot frameworks and platforms used. The elements for the conversational design of chatbots, as well as the terms used to describe them, vary between these frameworks and platforms.

In this case study, the framework Rasa [20] was used for chatbot development. Important basic elements are utterances, intents, entities, actions, and stories (e.g., [21]):

- *Utterances* are all expressions of users that are entered as user input into the chatbot user interface.
- *Intents* refer to goals that a user intends to achieve with the dialogue or information needs that a user wants to satisfy through communication with the chatbot.
- *Entities* modify or specify an intent and are extracted from the intent by the chatbot solution for further processing. This can be, for example, time and date information, places, names, quantities, etc.
- *Actions* define the output of the chatbot as a reaction to a certain intent and can contain not only text, but also links, buttons, graphical elements or videos. For natural language communication, however, text output is the main focus.
- *Stories* are used to link the different elements with each other, e.g., to specify a defined action for a certain intent.

According to [13] and [16], it can be distinguished between one-shot questions and those allowing for subsequent follow-up inquiries: In the simplest case or with one-shot queries from users, the chatbot generates a specific answer to a specific question (e.g., user: *"What university degree do I need for the job?"* → chatbot: *"You need a master's degree in electrical engineering."*). However, more complex dialogues are only possible if the context of successive questions is taken into account and, for example, more advanced follow-up queries are possible (e.g., user: *"Do I need a university degree for the job?"* → chatbot: *"Yes, a master degree."* → user: *"In which subject?"* → chatbot: *"In electrical engineering."*). This paper is a work-in-progress and initially deals with a chatbot prototype that focuses on successive one-shot queries and thus abstracts from the complexity of a contextual dialogue. In the following, the focus therefore lays on the identification and analysis of intents. However, for a more natural dialogue, aspects of the context must be taken into account in the future if the chatbot is developed further.

B. AI-Based Chatbot Implementation and Training Measures

Over the years, several different attempts proved valuable to create an AI which is able to respond to human queries and can thus be used as a foundation for advanced chatbots. It is possible to use sequence to sequence models [22][23]. For that purpose, the encoder processes the incoming query and generates a vector representation of the query. Queries contain a certain intent. As mentioned before, an intent expresses the user's intention he pursues with a made query in the sense of completing a certain task, for example to find a specific information [24]. Intents can thus be defined as predefined classes incoming inquiries can be categorised into and represent the types of queries the chatbot is capable of handling [7][25]. The decoder uses the established query representation to generate an answer. As a benefit, there is no need for a distinct set of answers, but answers are completely generated based on the user's input. In general, the models used for this do not need a task-specific setup; a domain specific corpus is required which contains generic queries and answers. But such corpora are quite scarce and rarely freely accessible.

Another possibility is to generate a vector representation of the incoming query and compare that representation to the ones of already known queries trying to find the best match [26]. If a reasonable match is found, it will be assumed that the new query is about the same intent as the known one or if no reasonable match is found, it can be assumed to have encountered an unknown query. It aims at clustering incoming queries and assign a general answer to each cluster. Although there are no unique answers created as for the sequence to sequence modeling, it is possible to easily expand the scope of such a system by adding new answers to the algorithm. The main problem is that generating sentence representation [27] is still challenging handling negations, contradictions and reciprocations.

Additionally, such a task can be seen as a classification problem. This completely limits the scope of the AI to the *a priori* set of answers. But the reduction in flexibility at least is accompanied with the AI model being specifically designed for the task [28]. In addition, the AI should be able to detect phrases, which are out of domain [29].

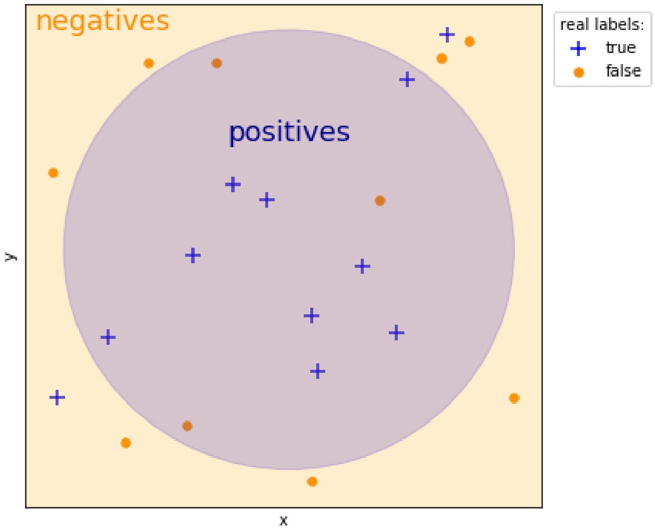


Figure 1: Example for the predictions of an algorithm

Figure 1 shows a possible prediction of an algorithm. All data points within the circle are predicted as true by the algorithm, the data points outside of the circle are predicted as false. A true positive is a data point that has the label "true" and the algorithm has also predicted this label for the data point. So a false positive actually has the label "false" but was predicted as true. The system is applied to the false labels. A true negative is a data point with the actual label false, which was also predicted by the algorithm as false. Finally, a false negative is a data point which was predicted as being true but with the real label "false". In the example above (Figure 1) there are eleven data points with the label "true" and ten with the label "false". An algorithm tried to predict the labels and predicted all data points within the circle as being true. This results in nine true positives and three false positives. An important measure for the accuracy in intent classification is the F1-score F_1 (e.g., to be seen in [30]). It is the harmonic mean of precision p and recall r. The precision p denotes the share of true positives from all positives. So in the example above (Figure 1), there are twelve data points determined as being true (the positives), but only nine of them are actually true (the true positives). Therefore, the precision $p = \frac{9}{12} = 0.75$. The recall denotes the share of true positives from all true labels. There are nine true positives in the example, but overall there are eleven data points with the label true. So, the recall results in $r = \frac{9}{11} \approx 0.82$. The F1-score is the finally calculated by $F_1 = 2 \cdot \frac{p \cdot r}{p+r} \approx 2 \cdot \frac{0.75 \cdot 0.82}{0.75+0.82} \approx 0.78$. For a non-binary classification problem, there is an F1-score for every label and the overall F1-score is usually calculated by averaging.

C. *FAQ-Chatbots in Recruiting*

This study applies an AI algorithm to the case of recruiting chatbots. As introduced, the recruiting process is especially suitable for efficiency enhancement by automation technology implementation [9]. Chatbots as automated dialogue systems can be deployed in various steps of the recruiting process to unburden the recruiters and leave them with the more strategic parts of the work while increasing efficiency as well as to reduce costs. They comply with the newly established requirements of potential candidates, who demand digital touch points in the form of mobile accessible websites and instant messaging [31]–[33]. According to a recent study in North America and Europe by Spiceworks [34], among organizations that currently deploy chatbots, 23 percent of administrative departments are equipped with such dialogue systems and seven percent already utilize this technology specifically within their human resource departments. Areas of application for chatbots along the recruiting process, some of which requiring components of artificial intelligence, are creating and posting job profiles, assisting job searches and the specific application process of potential candidates, handling incoming queries by applicants concerning general questions, support of recruiters during candidate pre-selection as well as during the hiring process [7]. Through automation, the efficiency of conducting these steps is improved [9]. Furthermore, the employer attractiveness is enhanced through chatbot implementation: According to a study by Phenom People based on more than 20 million chatbot interactions across over 100 chatbot deployments of the company, the number of job seekers turning into candidates applying for the job almost doubles (increase from 12 percent to 23 percent when implementing a chatbot on the career website) and the amount of candidates completing an application increases by 40 percent [35]. However, chatbots need to be integrated into the recruiters' Applicant Tracking

Systems (ATS), which handle application data and recruiting workflows in order to realize these potentials and to enhance the recruiting process [7]. Furthermore, chatbots cannot be seen as solution for any kind of application area and are no solution for all problems potentially occuring in recruiting.

The creation and integration of AI-based chatbots into ATS systems is being regarded in the governmentally funded research project CATS (Chatbots in Applicant Tracking Systems), which is a conjoint initiative of RheinMain University of Applied Sciences in Wiesbaden, and the talent management company milch&zucker AG in Gießen, Germany. This study project aims at the creation of a conceptual framework for a flexibly configurable chatbot toolbox, which is implementable prior, during and after the application process. An assortment of appropriate use cases as well as suitable intents is essential for relevant chatbot development. For specific use case selection of this study, interviews and surveys have been conducted with (1) technical, (2) scientific, and (3) industry experts concerning recruiting. The participants agreed upon FAQ scenarios (process guidance, application- and workflow-related questions, and guidance through the onboarding process) to be the most relevant and realistic in terms of support by chatbot implementation. This result is consistent with industry observations, which found questions related to the application status, job search and the company itself to be most common for chatbot inquiries [35]. Hence, these FAQ-related scenarios have been implemented into this study's case and an item set for an FAQ recruiting chatbot complying with these content requirements has been created.

III. Related Work and Research Objectives

A. Literature Review

Several studies already investigated the effects of AI in general (e.g., [9][10][36]) and chatbots in particular (e.g., [37]–[39]) on the recruiting process. However, as opposed to many studies incorporating either (1) perfunctory intent creation descriptions neglecting a comprehensive discussion of imperative underlying strategic considerations (e.g., [40]–[43]), (2) proposals of evaluation, i.e., rating and training methods for diverse chatbot prototypes without disclosure of the intent creation process (e.g., [44]–[47]), or (3) general investigation of intent matching and classification methods only (e.g., [48]–[51]), the interplay of intent creation and intent analysis within conversational design is not well covered by scientific research. Only two studies were found that deal with both the creation and the evaluation of intents for the use cases of (1) a hotel assistant chatbot [52] and (2) a Latvian customer support chatbot [53].

The most common error encountered within chatbot deployment according to the aforementioned study by Spiceworks [34] is the misunderstanding of incoming queries. This can refer either to (1) the intent matching capabilities of the chatbot framework, or (2) to the underlying intent set itself. Hence, developing and refining the most suitable list of intents alongside matching training and test data is fundamental to successful chatbot deployment. Encompassing evaluation is another crucial part of dialogue system design [15][46]. Human assessment is necessary within the evaluation of chatbots, either to (1) measure absolute task success, or (2) investigate user satisfaction on a more fine grained scale [4]. According to Walker et al., the users' perception of task completion success can predict user satisfaction better than actual task completion success [54]. Thus, an evaluation of the chatbot prototype from the users' perspective is conducted in this study with four users via rating of its response quality prior and after training (see Section IV-B). This kind of analysis is defined as session level user satisfaction evaluation [46].

B. Research Gap and Objective

There is an apparent lack of encompassing research dealing with both the establishment and the iterative adjustment process based on the evaluation of suitable chatbot intent sets. As seen throughout the literature review, only very few studies are known to the authors that disclose an in-depth approach to intent creation and evaluation through pre- and post-training tests of the different versions at the same time. This study gives detailed insights to the process of intent set creation and enhancement and furthermore proposes a structured approach for a recruiting FAQ chatbot. The central research questions are:

1) What is a relevant intent set for an FAQ recruiting chatbot?
2) Which effects can be seen when training the chatbot with enhanced data (intents and formulation variations) for improvement?

In the following, the approach to answer these research questions will be explained within the methodological section of the study.

IV. Methodology and Case Study Approach

As shown in the literature review, the identification of user intentions is an essential starting point of user-centric conversational design for chatbots. In the following, a case study in recruiting is used to describe how a basic set of user intents can be generated from various information sources. Starting from this basic set, the intents are analysed, cleaned up and provided with variations of user queries in a multi-stage process involving users. In addition, AI models are trained and evaluated with the sets of intents and corresponding variations of user questions. Finally, the resulting AI-based chatbots versions are subjected to user tests in order to evaluate the achieved quality of intent recognition.

A. User-centered Intent Identification

As described in the introduction, the starting point of our case study presented in this paper is first of all the composition and structuring of a comprehensive list of intents in the context of recruiting. Therefore, this section will explain the methodological approach in the sense of a user-centered attempt to identify user intents towards the chatbot (user centered intent identification) as well as suitable alternative formulations (to be used later as training and test questions). The following two sections then describe the approaches used to analyze and consolidate the developed intent sets, as well as the effects of modifications and model training on selected metrics and satisfaction values when using a corresponding chatbot prototype.

The overall methodological approach consists of five steps:

1) *Intent Sourcing*: Accumulation of potential intents from (1) website FAQs, (2) mail inquiries, (3) an expert review, and (4) user tests (see Figure 2).

2) *Intent Funneling*: Reduction of the initial item set via consolidation, reviewing and merging processes.
3) *Intent Variation*: Variation of the finalized item set through word substitution and splitting in into training and testing phrases.
4) *Intent Optimization*: Optimization of the item set through training, testing and intent matching coefficient improvements.
5) *Intent Validation*: The finalized item set is validated via a structured user test.

From the four sources described within the intent sourcing process, almost 500 initial intent propositions were drawn, which were reviewed, merged and eliminated in case of duplicates so that 82 final items emerged (see funneling process in Figure 2).

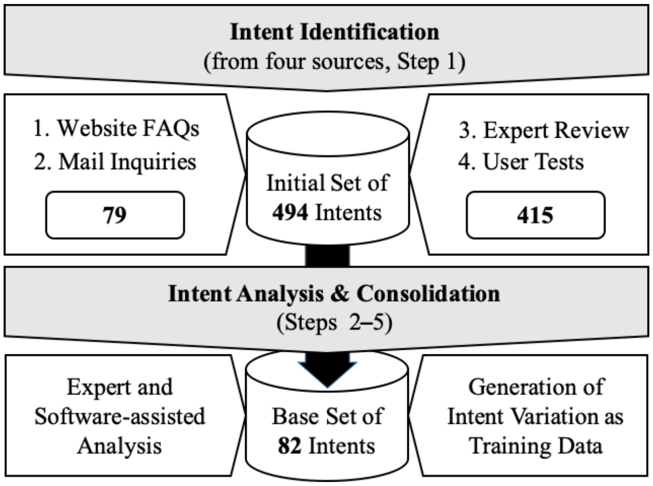


Figure 2: Overview of Intent Identification and Analysis

B. Analysis and Consolidation of Intents

After creation of the data set (base set), several instances of a natural language understanding artificial intelligence were trained to classify the intent of an input phrase. As mentioned before, the framework Rasa was used in general for the AI. Five different pipelines for the processing of the input messages were created and the DIET classifier was used in all instances. Sparse features were created in all cases by count vectorization of n-grams. The first one consisted of a white space tokenizer and only created sparse features for the tokens by the means of a Regex featurizer and count vectorization of words and n-grams. The second one additionally used the spacy components for creating tokens and dense features. The third one used the HFTransformerNLP with the ”Bert”-Model applying the bert-base model-weights for uncased words implementing the associated tokenizer and featurizer. In the remaing two instances, a white space tokenizer was used again and a neural network incorporating a biLSTM was used to create dense word embeddings from the char sequences of the input words. The corpus used to train both of these networks was chosen specifically for the task of job search consisting of over 400,000 job ads and 12,000 anonymized support emails from a company’s human resources management. One of these networks was trained by the approach by Ling *et al.* [55], training the previously mentioned embedding network as a part of a natural language processing task. In the other one, the network was trained to mimic the vectors created by a glove embedding as suggested by Pinter *et al.* [56]. These two networks based on character sequences rather than look-up tables were chosen in order to prevent the occurrence of out-of-vocabulary (oov) words. For comparison of all these setups, a five-fold cross-validation was performed on all models. The instance using the character to word embedding network as suggested by Ling *et al.* outperformed the other instances reaching an F1-score of 0.81 in average (see figure 3). The second best setup was the one incorporating the word embedding model of Pinter *et al.* with an average F1-score of 0.80. The spacy setup followed in third place barely beating the pure count vectorization of words approach by 0.01 comparing their respective F1-scores of 0.78 and 0.77. The most plausible explanation for the character-based neural networks outperforming spacy is that the corpus for them was chosen specifically for the task, while a general corpus based on news articles was used for spacy. Surprisingly, although consisting of a very sophisticated architecture, the Bert model performed worst for this task reaching an F1-score of only 0.69 in average. Fine-tuning the Bert-model might drastically improve the performance, but as Bert is by far the most resource-consuming model in this study and also performing worst, it was further excluded from investigation.

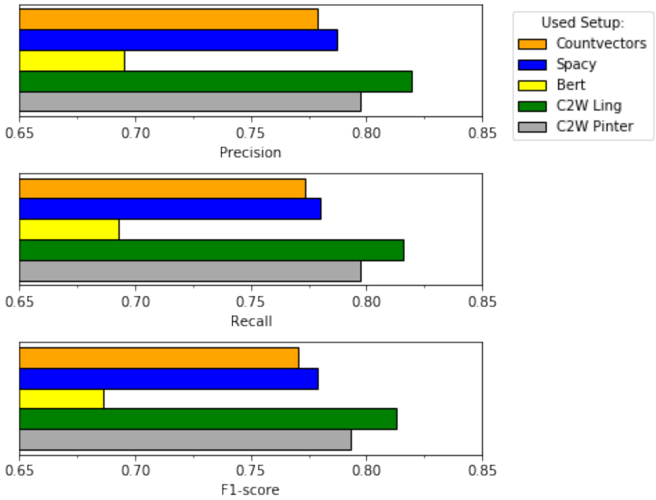


Figure 3: Comparison of setups for Rasa

To understand the sources of the errors, the confusion matrices were investigated and compared to the cosine similarities between the phrases of all intents. To calculate the cosine similarities, a sparse vector was assigned to every phrase with every entry consisting of the text frequency inverted document frequency value for every word in the corpus. This allowed the detection of several nearly or fully identical phrases within different intents, which explained at least some of the errors in the intent classification task. The data set was reworked, eight of the intents were removed and the corresponding phrases were shifted to other intents, ten were reworked and two new ones were created. The set of answers was reworked, too in order to fit to the new list of intents. Again, a five-fold cross-validation was performed to estimate the performance. This time, only the previously best-performing pipeline was used applying the character sequence to word embedding model of

Ling *et al.*

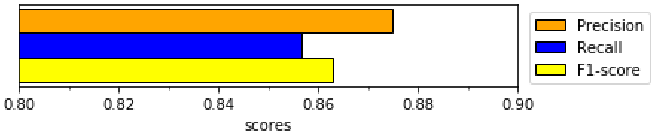


Figure 4: Scores for the setup using the character-based word model by Ling *et al.* after reworking the corpus

The F1-score was 0.86 in average for the new dataset (see Figure 4). A comparison of the scores of the two data sets was neglected, as there is a different number of classes and data points used. A reduction in classes should in general be accompanied with an increase in accuracy.

The intra-rater reliability $\kappa = \frac{p_0 - p_c}{1 - p_c}$, where p_0 is the accuracy of the chatbot in choosing the right intent and p_c is the probability to select the right intent by chance, is a measure showing how reliably a query is classified to the right intent. This reliability metric was calculated for both chatbot instances, the one trained on 88 and the one on 83 intents, to be 0.81 and 0.85 respectively (see Figure 5).

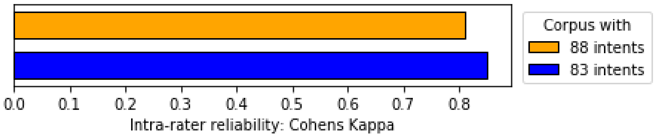


Figure 5: Intra-rater reliability for the two corpora with a different number of intents

It is important to note that the two values cannot be directly compared but can provide qualitative measure of performance. Although this metric does also not allow a direct comparison, a value above 0.8 usually shows that the predictions made by an algorithm are substantially reliable and not caused by chance. The higher score of 0.85 for the refined version might suggest also an higher reliability whereas viewing this as a general improvement has to be done with great care.

C. Measuring the Impact of Improved Intent Sets

To still compare the two variants of the chatbot, it was tried to capture the user experience when confronted with the AIs. In order to do so, two new instances of the chatbot using the well performing char sequence to word vector embedding were trained on their respective whole data set. One data set being the original one and the other data set being the new one with an reduced number of intents and reformulated answers. An independent test set consisting of 1,400 phrases was created and both versions of the chatbot predicted the answers to these phrases. Finally, the number of four students raters, R1 to R4, had to rate these answers as ”good”, ”mediocre” or ”bad”. They were asked to rate an answer as ”good” if the answer fitted the question. A ”mediocre” answer meant, that the chatbot gave an answer which at least corresponded to the right topic but did not exactly answer the question. A ”bad” answer was one that did not match the intent at all. This threefold evaluation scheme is loosely based on [44], who rated the appropriateness of the dialogue system based on the three categories (1) appropriate, (2) interpretable (evasive answer), and (3) inappropriate.

The students rated the two chatbots quite differently: One of the raters strongly favored the chatbot training on the refined corpus, giving ”good” ratings more often while reducing the number of ”mediocre” and ”bad” ratings for its answers. Two of the student raters only favored the refined version of the AI by a slight margin, with the tendencies towards ”good” ratings being less distinct as compared to the first student. The remaining student even gave fewer ”good” ratings for the answers of the refined version of the chatbot. Also, this student rated fewer answers as ”bad” mainly resulting in an increase in ”mediocre” answers. Overall, the answers of the refined chatbot were rated ”good” and ”mediocre” more often (72.0% vs. 71.1% and 7.5% vs. 6.8%) and ”bad” less often (20.5% vs. 22.1%) (see Figure 6). Due to the low number of test persons, especially viewing the standard deviation of the ratings, these results are not significant enough to claim a general trend.

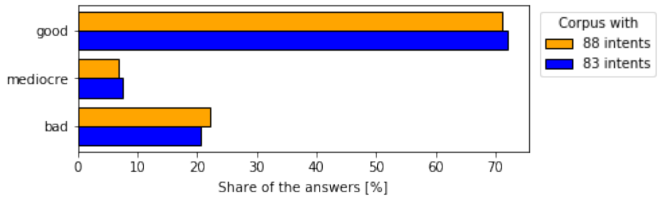


Figure 6: Ratings of the answers in average over all queries and students

One major problem of such a small number of test persons is that different mindsets are not averaged out and strongly dictate the outcome of the testing. Hence, all answers from both chatbot setups (corpora) were picked where all students gave the same rating. These should be very ”good” or very ”bad” answers, as the idea of what is ”mediocre” is more a question of the mindset than the extremes which are ”very good” or ”very bad”. Unsurprisingly, no answer was rated ”mediocre” by all students, but 67.4% of the answers for the chatbot trained on the original corpus with 88 intents and 69.6% of the answers of the refined chatbot got the same rating (see Figure 7).

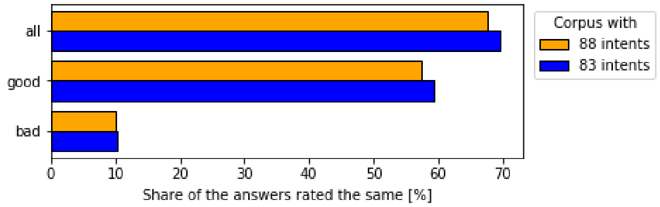
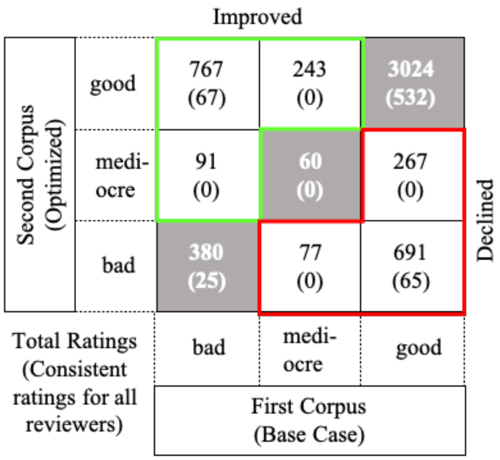


Figure 7: Percentages of all queries where the students gave the same rating for the answer of the corresponding chatbot

For the original chatbot 10.1% and for the refined one 10.2% of the answers were rated ”bad”. One might suggest that the the general setup of the chatbot combined with the limited training data is just not capable of understanding queries that are unknown. So, either the training corpus has to be extended or the word embedding needs to better capture semantic similarities. Further, 57.4% of the answers for the first version and 59.4% of the answers for the refined version were rated as ”good” by all test persons. This slight improvement at least suggests some positive effect of the intent refinement.

Focusing on consistent ratings, it can be seen in Figure 8 that out of the 6,500 evaluated cases in total, 3,464 ratings remained unchanged with either a consistant ”good”, ”mediocre” or ”bad” rating after the training of the chatbot. In 532 cases, all reviewers consistently gave a ”good” rating prior and after the training while the total amount of unchanged good ratings was 3,024. As mentioned before, there was no case of uniform ”mediocre” rating throughout the evaluation study amongst all four reviewers while the total amount of consistent ”mediocre” ratings was 60 out of the 5,600. 380 cases were rated badly in total while only 25 of them were reviewed as ”bad” by all four reviewers. Looking at the positive (improvement, edged in green) and negative changes (deterioration, framed in red), it becomes apparent that overall, more cases improved (1,101) than worsened (1,035) throughout the training.



Improved

Second Corpus (Optimized)	good	767 (67)	243 (0)	3024 (532)
	medi-ocre	91 (0)	60 (0)	267 (0)
	bad	380 (25)	77 (0)	691 (65)
Total Ratings (Consistent ratings for all reviewers)		bad	medi-ocre	good
		First Corpus (Base Case)		

Declined

Figure 8: Overview of the user rating distribution

In Figure 9, the rating changes through the training of the chatbot are broken further down. While several of the reviewers’ ratings seem to be similar, there are some noticeable differences between R1 and R3, especially regarding the verbatim ratings and the decline from the first towards the second corpus of the chatbot.

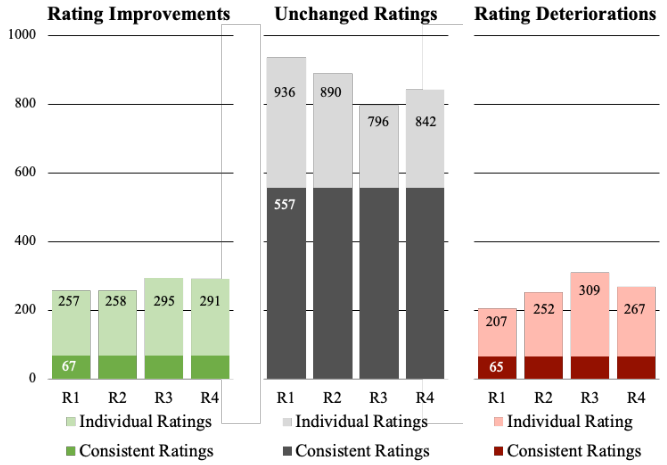


Figure 9: Rating comparison of improvement, verbatim state and deterioration

For the unchanged rating amounts, there is a spread of 140 differently rated cases and regarding the deteriorations, a gap of 98 stands out. However, with an exception of R3, the improvements or unchanged ratings outweigh the potential deterioration of the rating structure.

V. Conclusions and Managerial Implications

In summary, the chatbot composition and especially its conversational design is not finished yet. The training corpus still seems to be too narrow, suggested by the number of answers rated as ”bad” by all test persons. On the other hand, it seems that a useful setup for the embedding was found and that the refinement of intents had some effect. In the end, a lot of minor improvements will give rise to an overall powerful chatbot system.

The use of chatbots in recruiting will play a prominent role in the next few years in the handling of service dialogues within the organisation and towards the candidate. Especially in companies and organizations with a high number of applicants, the support of candidates in the recruiting process takes a considerable amount of time with frequently recurring requests for the same information. Here lies a significant savings potential on workforce (man-days) through the use of chatbot offers without worsening the quality of support. Furthermore, chatbots will play an increasingly important role in the dialogue between the hiring mamanger and the personnel/recruiting department. Chatbots will help with general FAQs but also with questions about the requirements of positions (skill management) or about the classification according to collective bargaining agreements, and they will also help with the formulation of advertisements. As they mature, they will also be able to help in the selection of applicants by autonomously conducting structured interviews.

The most important component to get there is to provide the best possible recognition of intents in the respective specific domain. As a basis, the conversational design in the form of a relevant and suitable intent set is indispensable. General user acceptance will then depend largely on apt responses and thus relevant content as well as a low number of incorrect answers within the dialogues.

VI. Limitations and Further Research

This case study described how an initial intent set for a FAQ chatbot can be developed for a specific application in recruiting and enhanced via a structured consolidation process (see Figure 2). The conversational design of this chatbot was initially limited to single-shot queries. Follow-up queries or a more complex context has not yet been considered in the dialogue modeling. In the following research work, such follow-up queries and context must also be included in dialogue modelling. It should also be noted that a user-centered improvement of the chatbot prototype can only be achieved if it interacts more extensively with real users and the resulting questions are used to extend and improve intent recognition. Nevertheless, the limits of the current chatbot development have already become clear due to the simple design. Only if the relevant intents can be captured, the chatbot will be able to provide a real benefit. The case study also showed that interdisciplinary cooperation between experts from different fields is necessary to successfully develop a chatbot. Conversational designers need to understand the basics of interactive design for conversational interfaces as well as the basics of AI solutions. Further research should also focus on how teams should be put together and which specific qualifications and skills are required for the individual roles and phases of the chatbot development process. Furthermore, a larger set of participants need to be exposed to the chatbot as a next step in order to yield generalizable information.

A research gap is also evident in the area of how the technical quality of an AI model and its improvement is related to the effect on the users. Developers and operators of chatbot solutions need, for example, technically derived information on how much training data is required or how a set of intents can be suitably improved. This is imperative as user tests are often complex and expensive and can have performance deficits at various levels of conversational design and AI components. Better research of such correlations is the basis for more sound recommendations for the design of chat offers in practice.

Acknowledgment

The study was carried out as part of the research project CATS (Chatbots in Applicant Tracking Systems) of the Rhein-Main University of Applied Sciences. This project (HA project no. 642/18-65) is funded in the framework of Hessen Modell-Projekte, financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

References

[1] J. Rowley, “Just another channel? Marketing communication in e-business,” *Marketing Intelligence & Planning*, vol. 17, no. 4, pp. 332–351, 2006.

[2] S. Böhm and J. Eißer, “Hedonic motivation of chatbot usage: Wizard-of-oz study based on face analysis and user self-assessment,” in *Proceedings of CENTRIC 2017*, 2017, pp. 59–66.

[3] A. Mittal, A. Agrawal, A. Chouksey, R. Shriwas, and S. Agrawal, “A comparative study of chatbots and humans,” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6, no. 6, pp. 1055–1057, 2016.

[4] D. Jurafsky and J. H. Martin, “Speech and language processing (draft),” *Chapter 24: Dialog Systems and Chatbots (Draft of September 11, 2018). Retrieved March*, vol. 19, p. 2019, 2018.

[5] Research and Markets, *Chatbot market by component (solutions and services), usage (websites and contact centers), technology, deployment model, application (customer support and personal assistant), organization size, vertical, and region - global forecast to 2024*, 2019. [Online]. Available: https://www.researchandmarkets.com/reports/4858082/chatbot-market-by-component-solutions-and? [retrieved: 07/13/2020].

[6] L. Schildknecht, J. Eißer, and S. Böhm, “Motivators and barriers of chatbot usage in recruiting: An empirical study on the job candidates perspective in germany,” *Journal of E-Technology Volume*, vol. 9, no. 4, pp. 109–123, 2018.

[7] S. Meurer, J. Eißer, and S. Böhm, “Chatbots in applicant tracking systems: Preliminary findings on application scenarios and a functional prototype,” in *Proceedings of the IWEMB 2019: Third International Workshop on Entrepreneurship in Electronic and Mobile Business*, in press.

[8] G. V. Research, *Chatbot market size, share & trends analysis report by end user, by application/business model, by type, by product landscape, by vertical, by region, and segment forecasts, 2018 - 2025*, 2017. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/chatbot-market [retrieved: 07/13/2020].

[9] B. Hmoud *et al.*, “Will artificial intelligence take over human resources recruitment and selection,” *Network Intelligence Studies*, vol. 7, no. 13, pp. 21–30, 2019.

[10] S. İşgüzar and C. Ayden, “New decision mechanism in the recruitment process: Artificial intelligence,” in *A New Perspective in Social Sciences*, London: Frontpage Publications, pp. 196–205.

[11] S. Reviews, *The Top 10 Best Recruiting and HR Chatbots - June 2020*, 2020. [Online]. Available: https://www.selectsoftwarereviews.com/buyer-guide/hr-chat-bots [retrieved: 07/12/2020].

[12] A. Multiple, *Top 60 Chatbot Companies of 2020: In-depth Guide*, 2020. [Online]. Available: https://research.aimultiple.com/chatbot-companies/ [retrieved: 07/12/2020].

[13] M. F. McTear, “The rise of the conversational interface: A new kid on the block?” In *International Workshop on Future and Emerging Trends in Language Technology*, Springer, 2016, pp. 38–49.

[14] C. Pricilla, D. P. Lestari, and D. Dharma, “Designing interaction for chatbot-based conversational commerce with user-centered design,” in *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 2018, pp. 244–249.

[15] M. McTear, “Conversation modelling for chatbots: Current approaches and future directions,” *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, pp. 175–185, 2018.

[16] S. Janarthanam, *Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Birmingham: Packt Publishing, 2017.

[17] J. Feine, S. Morana, and U. Gnewuch, “Measuring service encounter satisfaction with customer service chatbots using sentiment analysis,” in *14th International Conference on Wirtschaftsinformatik*, 2019, pp. 1115–1129.

[18] R. Batish, *Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger*. Birmingham: Packt Publishing, 2018.

[19] A. Fadhil, “Domain specific design patterns: Designing for conversational user interfaces,” *arXiv preprint arXiv:1802.09055*, 2018. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1802/1802.09055.pdf [retrieved: 07/12/2020].

[20] Rasa, *Rasa: Open source conversational AI - Rasa*, 2020. [Online]. Available: https://rasa.com/ [retrieved: 07/12/2020].

[21] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, “Rasa: Open source language understanding and dialogue management,” *arXiv preprint arXiv:1712.05181*, 2017. [Online]. Available: https://arxiv.org/abs/1712.05181 [retrieved: 07/12/2020].

[22] O. Vinyals and Q. Le, *A neural conversational model*, 2015. arXiv: 1506.05869. [Online]. Available: https://arxiv.org/abs/1506.05869 [retrieved: 07/12/2020].

[23] A. Sojasingarayar, *Seq2seq ai chatbot with attention mechanism*, 2020. arXiv: 2006.02767. [Online]. Available: https://arxiv.org/abs/2006.02767 [retrieved: 07/12/2020].

[24] U. Şimşek and D. Fensel, “Intent generation for goal-oriented dialogue systems based on schema. org annotations,” in *1st International Workshop on Chatbots Co-Located with the 12th International Conference on Web and Social Media (ICWSM2018)*, 2018, pp. 1–7.

[25] S. Srivastava and T. Prabhakar, “Intent sets: Architectural choices for building practical chatbots,” in *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering*, 2020, pp. 194–199.

[26] N. Lair, C. Delgrange, D. Mugisha, J.-M. Dussoux, P.-Y. Oudeyer, and P. F. Dominey, “User-in-the-loop adaptive intent detection for instructable digital assistant,” *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Mar. 2020.

[27] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, 2019. arXiv: 1908.10084. [Online]. Available: https://arxiv.org/abs/1908.10084 [retrieved: 07/12/2020].

[28] A. Perevalov, D. Kurushin, R. Faizrakhmanov, and F. Khabibrakhmanova, “Question embeddings based on shannon entropy: Solving intent classification task in goal-oriented dialogue system,” in *Proceedings of International Conference on Applied Innovation in IT*, 2019, pp. 73–78.

[29] S. Larson *et al.*, “An evaluation dataset for intent classification and out-of-scope prediction,” in *roceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1311–1316.

[30] B. Liu and I. Lane, “Attention-based recurrent neural network models for joint intent detection and slot filling,” in *Interspeech 2016*, 2016, pp. 685–689. DOI: 10.21437/Interspeech.2016-1352. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1352.

[31] C. Lieske, “Digitalisierung im Bereich Human Resources (Digitization in the field of human resources),” in *Digitalisierung in Industrie-, Handels-und Dienstleistungsunternehmen (Digitization in industrial, commercial and service enterprises)*, Wiesbaden: Springer Gabler, 2020, pp. 149–160.

[32] D. Bollessen, *Der fortschreitende Fachkräftemangel infolge des demographischen Wandels: Denkbare Konzepte und Erfolgsstrategien zur langfristigen Mitarbeiterbindung (The continuing shortage of skilled workers as a result of demographic change: Conceivable concepts and successful strategies for long-term employee retention)*. Diplomica Verlag, 2014.

[33] R. Hartmann, “Rekrutierung im Mittelstand: Trends und Herausforderungen im Personalmanagement oder von Trüffelschweinen und Wollmilchsäuen (Recruitment in medium-sized businesses: trends and challenges in human resources management or of truffle pigs and Jack of all trades),” in *Rekrutierung in einer zukunftsorientierten Arbeitswelt (Recruiting in a future oriented working world)*, M. Hartmann, Ed., Wiesbaden: Springer Gabler, 2015, pp. 215–234.

[34] Spiceworks, *Data Snapshot: AI Chatbots and Intelligent Assistants in the Workplace*, 2018. [Online]. Available: https://community.spiceworks.com/blog/2964-data-snapshot-ai-chatbots-and-intelligent-assistants-in-the-workplace [retrieved: 07/13/2020].

[35] Phenom People, *Chatbots for Recruiting - Benchmarks 2020*, 2020. [Online]. Available: https://www.phenom.com/resource/chatbots-for-recruiting-2020-benchmarks [retrieved: 07/13/2020].

[36] Q. Jia, Y. Guo, R. Li, Y. Li, and Y. Chen, “A conceptual artificial intelligence application framework in human resource management,” in *Proceedings of the International Conference on Electronic Business*, 2018, pp. 106–114.

[37] Q. V. Liao *et al.*, “All work and no play?” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[38] N. Nawaz and A. M. Gomes, “Artificial intelligence chatbots are new recruiters,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, pp. 1–5, 2019.

[39] G. Suciu, A. Pasat, C. Bălăceanu, C. Nădrag, and A. Drosu, “Design of an internship recruitment platform employing nlp based technologies,” in *ECAI 2018-International Conference, June 2018*, 2018, pp. 1–6.

[40] M. Fleming *et al.*, “Streamlining student course requests using chatbots,” in *29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018)*, Engineers Australia, 2018, pp. 207–211.

[41] G. M. Mostaco, I. R. C. De Souza, L. B. Campos, and C. E. Cugnasca, “Agronomobot: A smart answering chatbot applied to agricultural sensor networks,” in *14th international conference on precision agriculture*, vol. 24, 2018, pp. 1–13.

[42] M. Carisi, A. Albarelli, and F. L. Luccio, “Design and implementation of an airport chatbot,” in *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, 2019, pp. 49–54.

[43] T. L. Vu, K. Z. Tun, C. Eng-Siong, and R. E. Banchs, “Online faq chatbot for customer support,” in *Proceedings of the 2019 10th International Workshop on Spoken Dialogue Systems Technology*, 2019, pp. 1–6.

[44] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, “Chatbot evaluation and database expansion via crowdsourcing,” in *Proceedings of the chatbot workshop of LREC*, 2016, pp. 15–19.

[45] Å. Kamphaug, O.-C. Granmo, M. Goodwin, and V. I. Zadorozhny, “Towards open domain chatbots—a gru architecture for data driven conversations,” in *International Conference on Internet Science*, 2017, pp. 213–222.

[46] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, “A survey on evaluation methods for chatbots,” in *Proceedings of the 2019 7th International Conference on Information and Education Technology*, 2019, pp. 111–119.

[47] E. Ruane, R. Young, and A. Ventresque, “Training a chatbot with microsoft LUIS: Effect of intent imbalance on prediction accuracy,” in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 2020, pp. 63–64.

[48] B. Behera, “Chappie-a semi-automatic intelligent chatbot,” *Write-Up*, pp. 1–5, 2016.

[49] M. Y. H. Setyawan, R. M. Awangga, and S. R. Efendi, “Comparison of multinomial naive bayes algorithm and logistic regression for intent classification in chatbot,” in *2018 International Conference on Applied Engineering (ICAE)*, 2018, pp. 1–5.

[50] S. Alias, M. S. Sainin, T. S. Fun, and N. Daut, “Intent pattern discovery for academic chatbot-a comparison between n-gram model and frequent pattern-growth method,” in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2019, pp. 1–5.

[51] K. Balodis and D. Deksne, “Fasttext-based intent detection for inflected languages,” *Information*, vol. 10, no. 5, p. 161, 2019.

[52] L. N. Michaud, “Observations of a new chatbot: Drawing conclusions from early interactions with users,” *IT Professional*, vol. 20, no. 5, pp. 40–47, 2018.

[53] K. Muischnek and K. Müürisep, “Collection of resources and evaluation of customer support chatbot,” in *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018*, 2018, pp. 30–37.

[54] M. Walker, C. Kamm, and D. Litman, “Towards developing general models of usability with paradise,” *Natural Language Engineering*, vol. 6, no. 3 & 4, pp. 363–377, 2000.

[55] W. Ling *et al.*, “Finding function in form: Compositional character models for open vocabulary word representation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1520–1530.

[56] Y. Pinter, R. Guthrie, and J. Eisenstein, “Mimicking word embeddings using subword RNNs,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 102–112.