INTENT IDENTIFICATION AND ANALYSIS FOR USER-CENTERED CHATBOT DESIGN – A CASE STUDY ON THE EXAMPLE OF RECRUITING CHATBOTS IN GERMANY

Sebastian Meurer | Judith Drebert | Prof. Dr. Stephan Böhm
RheinMain University of Applied Sciences, Wiesbaden
Faculty Design – Computer Sciences – Media
Degree Program Media Management

Olena Linnyk | Jens Kohl | Harald Locke | Ingolf Teetz | Levitan Novakovskij
Milch & zucker AG
Gießen, Germany
01

INTRODUCTION
Chatbots are automated dialogue systems for conversational scenarios based on pattern matching or artificial intelligence (Mittal et al., 2016).

Such systems can automate dialogues between companies and customers for large scale utilization (Böhm & Eißer, 2017).

They hold a vast potential (Research & Markets, 2019):
- chatbot market worth 9.4 bn. USD by 2024
- 30% annual growth rate
Chatbot Use Cases
Chatbots Within the Recruiting Process

- Chatbots potentially support various business processes (Schildknecht et al., 2018; Meurer et al., 2020; G. V. Research, 2017)

- Especially feasible for FAQ scenarios (Hmoud & Laszlo et al., 2019)

- Increase efficiency while reducing costs (Hmoud & Laszlo et al., 2019) when applied in the company’s Applicant Tracking Systems (ATS)

- In recruiting, they can transfer information to potential candidates before, throughout and after the application process
  - Support within sourcing and screening processes
  - Reduction of human bias
  - Allows for recruiting activities at the most suitable points of contact for potential candidates; e.g. mobile accessible websites and instant messaging (Lieske, 2020; Bollessen, 2014; Hartmann, 2015)

- Recruiting chatbots are relatively new; solutions are often early test applications and not yet in permanent productive use

- Currently utilized by 7% of companies within HR (Spiceworks, 2018)
Motivation
Relevancy of Intent Definition Within Dialogue Creation

- HR decision makers sometimes think that chatbot solutions are autonomous learning systems building knowledge to answer user questions themselves.
- However, AI is limited to Natural Language Understanding (NLU) and question classification to predefined user intentions.
- These user intentions have to be created in the system and to be linked to certain actions for output.
- Hence, apart from technical implementation, chatbot developers need to define and structure dialogue contents in a conversational design (McTear, 2016).
- The intention selection is highly relevant: defines the application domain the chatbot can answer user requests in (Pricilla et al., 2018).
- Hardly any practical description of the intention selection procedure within literature (Pricilla et al., 2018).
Study Overview

Main Goals of the Study

This study

- describes necessity as well as the actual formation process of a suitable intent set for a corpus-based recruiting FAQ chatbot

- challenges a newly trained version of the chatbot against the former version of this dialogue technology prototype
02

RESEARCH BACKGROUND
(CATS – CHATBOTS IN APPLICANT TRACKING SYSTEMS)
Chatbots are **conversational interfaces** (McTear, 2018)

Special kind of interactive user interface: allows for **natural language dialogues** between humans and computers, oftentimes based on **AI functionalities** (McTear, 2016; Janarthanam, 2017)

Typically embedded in a **website** or **messaging solution** (Feine et al., 2019)

Conversational design is about **interface design** (e.g., stakeholder/goal definition, conversational flow design, development, testing) to **provide good user experience** (Janarthanam, 2017; Batish, 2018)

- Variations of colours, fonts or graphic elements (e.g., buttons, emojis)
- Personality
  - Tonality
  - **Dialogue content** and its logical structure as core of conversational design
  - **One-shot questions** vs. those allowing for subsequent **follow-up inquiries**
## Research Background (2)

### Conversational Design within RASA

**Conversational framework: RASA**

- Open source chatbot development platform

---

<table>
<thead>
<tr>
<th>Utterances</th>
<th>Intents</th>
<th>Entities</th>
<th>Actions</th>
<th>Stories</th>
</tr>
</thead>
<tbody>
<tr>
<td>All expressions of users that are entered as user input into the chatbot user interface</td>
<td>Goals that a user intends to achieve/information need users want to satisfy</td>
<td>Specification or modification of an intent</td>
<td>Define the output of the chatbot as reaction to an intent</td>
<td>Link different elements with each other</td>
</tr>
<tr>
<td><em>e.g., “I want to know how I can apply for the job XY.”</em></td>
<td>Predefined classes setting the capacity of the chatbot</td>
<td>Extracted from the intent for further processing</td>
<td>Can contain different kinds of elements</td>
<td>Specify a defined action for a certain intent</td>
</tr>
<tr>
<td></td>
<td><em>e.g., “Application procedure”</em></td>
<td><em>e.g., time, location, name, quantity</em></td>
<td><em>e.g., text, link, button, video</em></td>
<td><em>e.g., if intent “Application procedure”, action “Link to manual”</em></td>
</tr>
</tbody>
</table>

Sources: RASA, 2020; Bocklisch et al., 2017; Meurer et al., 2020; Srivastava & Prabhakar, 2020
Research Background (3)
AI-based Chatbot Implementation and Training Measures/Methods

✦ **Sequence to sequence models** (Vinyals & Le, 2015; Sojasingarayar, 2006)
  - Intents in the form of predefined classes and established query representation are utilized by the decoder to generate an answer
  - Hence, no distinct set of answers but generation based on user input
  - No task-specific setup but domain specific corpus (contains generic queries and answers)
  - Such corpora are scarce and rarely freely accessible

✦ **Vector representation** of incoming query and comparison of the representation to the ones of already known queries to find the best match (Lair et al., 2020)
  - In case of a reasonable match, it is assumed that the new query has the same intent as the known one
  - Incoming queries are clustered and general answers are assigned to each cluster
  - New answers have to be added to the algorithm
  - Problematic: sentence representation as the more words added, the more complex the matching process in terms of negations, contradictions and reciprocations (Neimers & Gurevych, 2019)
Research Background (3)

Limitations of AI, Accuracy Measurement with F1-Score

- Predefined answers result in an **AI-based a priori** set of answers
- Algorithm predictions can be visualized
  - Data points **within** circle: predicted as **true** by algorithm
  - Data p. **outside** circle: predicted as **false** by algorithm
  - Data: 11 labelled true; Algorithm: 9 → 2 false negatives
  - Data: 10 labelled false; Algorithm: 7 → 3 false positives
- Accuracy measurement in intent classification: F1-score $F_1$ (Liu & Lane, 2016)
  - $F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ (share of true positives of all predicted positives) and recall r (share of true positives from actually labelled positives)
  - In our example: $2 \times \frac{p \times r}{p + r} = 0.78$

Fig. 1: Exemplary algorithm prediction visualization

True positive: label true; algorithm predicted as true
False positive: label false; algorithm predicted as true
True negative: label false; algorithm predicted as false
False negative: label true; algorithm predicted as false
03 RELATED WORK AND RESEARCH OBJECTIVES
Related Work

Literature Review

🌟 Several studies investigated the effects of **AI in general** (Hmoud & Laszlo, 2019; Isgüzar & Ayden, 2019; Jia et al., 2018) and **chatbots in particular** (Liea et al., 2018; Nawaz & Gomes, 2019; Suciu et al., 2018)

🌟 **Interplay of intent creation and intent analysis** within conversational design not well covered by scientific research

🌟 Only **two studies** found dealing with the **creation as well as evaluation of intents** for (1) a hotel assistant chatbot (Michaud, 2018) and (1) a Latvian customer support chatbot (Muischnek & Müürisep, 2018)

🌟 However, the **misunderstanding** of incoming queries is the **most common chatbot error** (Spiceworks, 2018)

→ **Developing and refining** the most suitable list of **intents** is imperative

→ **Encompassing evaluation** as another crucial part of dialogue system design (McTear, 2018; Maroengsiet al., 2019)
Research Objectives

Apparent lack of encompassing research dealing with both the establishment and the iterative adjustment process based on the evaluation of suitable chatbot intent sets.

This study offers detailed insights to the process of intent set creation and enhancement.

Proposition of a structured approach for recruiting FAQ chatbot development.

Central research questions:

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?
04

METHODOLOGY AND CASE STUDY APPROACH
Methodology

General Approach

Approach:

(1) Intent generation from different information sources

(2) Intent analysis, cleaning and variation of intents

(3) Training and evaluation of the varied intents including user tests
Case Study Approach
User-centered Intent Identification

Five step approach:

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

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 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.

---

**Intent Identification**
(from four sources, Step 1)

1. Website FAQs
2. Mail Inquiries
3. Expert Review
4. User Tests

Initial Set of 494 Intents

- 1. Website FAQs: 79
- 2. Mail Inquiries: 415

**Intent Analysis & Consolidation**
(Steps 2–5)

- Expert and Software-assisted Analysis
- Generation of Intent Variation as Training Data

Base Set of 82 Intents

Fig. 2: Overview of Intent Identification and Analysis
Case Study Approach
Analysis and Consolidation of Intents

❖ Training in RASA (instances of NLU AI to classify the intents)
❖ Training corpus of 400,000 job ads and 12,000 anonymized support e-mails from companies’ human resources management
❖ Testing of different measures for best performance
❖ Best one: character to word embedding network as suggested by Ling et al. (2015)
  - *F₁* score of 0.81 on average
❖ Creation of confusion matrices to understand the sources of errors
❖ According rework of the data set:
  - 8 intents removed, phrases shifted to others
  - 10 intents reworked
  - 2 intents newly set up
  - Adaptation of the answer set
  - New *F₁* score of 0.86 on average
  - Predictions made by the algorithm substantially reliable and not caused by chance (intra-rater reliability of 0.85 as opposed to formerly 0.81 (no direct comparison possible but indication for improvement of reliability)
Case Study Approach

Measuring the Impact of Improved Intent Sets (1)

❌ For comparison of the two chatbot variants, the **user experience** was captured

❌ Old data set vs. new one (revised and reduced no. of intents, reformulated answers)

❌ Test approach:
  - Independent test set of 1,400 phrases
  - Algorithm of both chatbot versions predicted the answers
  - Loosely based on Yu et al. (2016), 4 student raters (R1-R4) rated the resulting answers as
    1) “good” (fitting answer),
    2) “mediocre” (answer of correct topic but no exact answer to the question), or
    3) “bad” (did not match intent at all)

❌ Refined chatbot (blue) yielded more positive (good & mediocre) ratings

❌ 57.4% of the answers for the old and 59.4% of the refined version were rated as ”good” by all raters

→ **positive effect**
Case Study Approach

Measuring the Impact of Improved Intent Sets (2)

- Out of 6,500 evaluated cases in total, 3,464 ratings maintained unchanged

- Unchanged good ratings: 3,024; in 532 cases, all reviewers consistently rated as “good”

- 380 cases rated badly; only 25 of those seen as “bad” by all 4 reviewers

- Overall, more cases improved (1,101) than worsened (1,035) throughout the training
The ratings across the four reviewers are noticeably different.

Overall, the improvements or unchanged ratings outweigh the potential deterioration of the rating structure.

Especially concerning the unchanged ratings, differences in between the raters become apparent.
05 CONCLUSION & MANAGERIAL APPROACH (LIMITATIONS, IMPLICATIONS FOR FUTURE RESEARCH)
Conclusion and Managerial Implications
Summary and Practical Implications

Conclusion

- Chatbot composition and especially the conversational design is a complex field
- The training as conducted in this study showed positive effects
- For the case at hand, the training corpus need some more revision/minor improvements
- Interdisciplinary cooperation between experts necessary to successfully develop a chatbot

Managerial Implications

- The use of chatbots in recruiting will play a prominent role in the next years
- Especially useful for companies with high volumes of applications
- Most important is correct intent recognition in the specific domain
- User acceptance will depend on apt responses and low numbers of incorrect answers
06 LIMITATIONS AND FURTHER RESEARCH
Limitations

- Student rater sample too small to yield significant, generalizable results
  - Problem: different mindsets are not averaged out and strongly dictate the outcome of the testing
- Single-shot queries regarded only (no context)

Outlook/Suggestions for future research

- Inclusion of follow-up queries into the research work
- User tests with the chatbot prototype itself
- Focus on how to form teams (qualifications and skills) for chatbot development process
- Retest with a larger set of participants to yield generalizable information
- Analysis of the relationship of the technical quality of an AI model with the users
THANK YOU!

DO YOU HAVE ANY QUESTIONS?

CENTRIC Paper Presentation 2020

Sebastian Meurer | Judith Drebert | Prof. Dr. Stephan Böhm
stephan.boehm@hs-rm.de
RheinMain University of Applied Sciences, Wiesbaden
Faculty Design – Computer Sciences – Media
Degree Program Media Management

Olena Linnyk | Jens Kohl | Harald Locke | Ingolf Teetz | Levitan Novakovskij
olena.linnyk@milchundzucker.de
Milch & Zucker AG
Gießen, Germany
Publicity regulation

CATS – Chatbots in Applicant Tracking Systems

This project (HA project no. 642/18-65) is funded in the framework of Hessen ModellProjekte, 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).

For more information: www.innovationsfoerderung-hessen.de
BACKUP & APPENDIX


Literature/Bibliography


in Electronic and Mobile Business, in press.