Wirtschaftsinformatik
und Maschinelles Lernen
Stiftung Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
Prof. Dr. Dr. Lars Schmidt-Thieme
Jonas Falkner, M.Sc.
Nils Westphal

Research Proposal

ChatBot a Conversational AI

Niraj Dev Pandey

271484
(pandey@uni-hildesheim.de)

26th November 2018

Abstract

Users are rapidly turning to company's customer care to receive customer support about their difficulties while using any products or services. However, most of these queries are not being addressed on time or in some cases not addressed at all. Which leads to customer dis-satisfaction. COMPRA GmbH is one of those companies, which is also witnessing this problem. With the increase in number of user base and services, the customer support which is being done via phone, e-mails and ticket systems gets unnecessarily time-consuming. Another problem is to have human assistants in customer support and the cost it causes to the company. One way to solve this problem is to create a new conversational system to automatically generate responses for customers requests via a chat-Bot. Currently, in domain specific conversational systems structured data is being used for the bot to train however, gathering such a data-set is not an easy task for many companies. Preparing training data for chat-bots in conventional format itself is a time consuming and costly process. This thesis present a solution for data structurization for chat-bot training where bot will be able to understand user inputs and answer them accordingly. The goal is to develop a chat-bot from unstructured data by building a novel framework for feature extraction from such data. First, the performance will be tested by human analyst within the COMPRA GmbH and then we will deploy the COMPRA bot into real world and evaluate it's performance on genuine customers' queries.

Contents

1 Introduction

Chatbots are computer programs that can communicate with humans in human understandable language. In the early days they were called *chat-robots.* The era of chat bot started in 1960's which were intended to see, if they can fool people that they are real human beings. Since last few years chat bots have became a buzz after the development of Apple's Siri, Microsoft's Cortona and Google's Google Assistant. Following sections discuss the various types of chat-bots and their characteristics.

1.1 Types of Chat-Bots

Chat-bot can be divided mainly into three types. Such as, *Question-Answering Agent*, *task-oriented dialogue system* and *Social* or *General purpose bot.*These categorization is based on the way the bot performs and the kind of tasks they tackle.The name itself is explanatory. Nevertheless, let's dive deeper to know types of Chat-Bot in detail.

1.1.1 Question-Answering Agent

Recent past has witnessed an increasing demand for question answering (QA) chat-bots that allow users to query large scale knowledge bases or document collections via natural language. Q-A agents are the simplest and straight forward bots. They can provide concise direct answers to user's questions. Most of the industries are using such type of bots for their business purposes. Most well known Q-A bots are being developed via platforms like *Dialogueflow.AI*, *Wit.AI,Amazon Lex* and *IBM Watson.* A Q-A bot works like a SQL system where you have a database and you query within the DB for any specific outcome. Here user query is the question and the outcome will be an answer. Later, people started calling it 'Knowledge Based- Question Answering system' or 'text-QA agent' [citeJianfengGao2018]. These bots are trained on question-answer conversation pairs, where the bot returns an *answer* for a query*(question)* . Since, there are several ways to ask for any information therefore, these bots have numerous question-answer pairs to answer one query. Below is an example of Ubuntu open source data-set which is helpful to understand data format of Q-A chat-bots.

Anybody familiar with laptop-mode here how come i m changing laptop-mode conf but nothing is actually changing you d need to reboot possibly what are you trying to change at first it seemed to work but when i rebooted it seems to do nothing ... i need to change the max cpu freq and hd idle time surely that s a powernowd config option powernowde works on intel cpus too *this is n't a p4 is it right but you have both installed if you want there s no real advantage to that tho ....and so on*

1.1.2 Task-oriented dialogue system

Task oriented chat-bots can also be called domain specific conversational systems. They can assist user from variety of tasks, such as fix a meeting, make a booking or look for a restaurant etc. These kind of bots can ask user for more information and return a response when all the *slots* are fulfilled. Slots are referred to as the crucial information that needs to be provided by the user before getting to the final answer. Such as, a restaurant bot will only reserve a seat or tell the name of the restaurant, once you have provided the *place, number of people* and other information alike. In case of task oriented chat-bots, data-set is always categorized into ’Intents’, ’Entities’ and ’Response’. Generally, a .json file format is preferred where we have a query followed by the intention and the entities. Response will be generated according to detected *intent*.

```
   {
”text”: ”moderately priced restaurant that serves creative food”,
”intent”: ”inform”,
”entities”: [
{
”start”: 41,
”end”: 49,
”value”: ”creative”,
”entity”: ”cuisine”
},
{
”start”:0,
”end”: 10,
”value”: ”moderate”,
”entity”: ”price”
}
]
},
```

The above data is the NLU part of the entire data-set used for task-oriented agents. Response generation is a mapping of intent detected and the answers found for that particular intent.

1.1.3 General purpose Chat-Bot

Here a person can talk to the bot as your companion. You can ask general questions which the bot is trained on. Such as, what's the color of the sky, I am sad today, would you like to drink a beer etc. These bots are mostly trained on publicly available data such as *Movie dialogues, Cornell, Opensubs, Scotus data* etc. These datasets contains conversations between two individuals or more. Since the conversation in dataset differs from context to context, general purpose bots tend to generate unexpected responses.

In industrial domain none of the customer or clients are interested to ask what's a color of the sky hence, our study is about domain specific chat-Bot development, which would be responsible to answer queries within a particular domain.

1.2 Motivation

Developing an intelligent dialogue system which is trained on unstructured data and can answer questions of topics ranging from one business module to other, has been one of the long awaited goal in AI. Since, manual creation of dataset takes resources that many companies are not ready to afford therefore, it is a big obstacle in the path of chatbot development. This study will address the preparation of structured data for chat-bot training and exploring possibilities to develop chatbot on unstructured text.

Another motivation behind this research is to address customer's queries economically. The amount of human workforce needed to answer costumers' queries are so much that companies are heading towards chat-bots and trying to use as less human call attenders as they can. Working hours is also one of the concern. Particularly Germany have 5 to 8 hours a day service centers open and only business days are applicable. What if user wants some assistant on weekend? He/she has to wait till next week which leads to customer dis-satisfaction in many cases. India is a leading call center service provider, one can find that Indian companies providing call service 24x7. Which results in huge economical challenges for companies. Some say "human sleeps, take vacations and salary machine doesn't". This is the prime reason for companies to automate tasks and services. Present research provides a solution to COMPRA GmbH to deal with customers' queries efficiently and economically.

1.3 Problem Statement

Collecting structured data for chat-bot development is a costly and time consuming process. Every industry documents their products and services providing information and FAQ's in some format. These documents are not useful for chat-bot development directly. Currently, manual data engineering techniques are required to convert the documents in a proper format. However, not all the businesses have resources to invest in such a costly and time consuming process. In other words, documents containing information about the products and services of a company are mostly available but needs manual techniques to further engineer them. This research aims to develop a framework which is able to give a structure to the raw data and can be used to train a chatbot model.

2 Literature Review

2.1 Research Questions

Many websites/companies have a domain specific chat-Bot on their web page to navigate user throughout their products and services. Most of them or in most cases all of them are developed on the structured data. Where either their data set is in the form of question and answer pair respectively or any other structure (Json, markdown etc). Collecting structured data for all the services and the products is a big concern as well as time consuming and an expensive process. This is also a reason that companies haven't been able to successfully develop a domain specific chat-Bot which can tackle a large scopic conversation.

One might have witnessed that for simple question bot replies "I am sorry! I am just a bot, go easy with me". This is the case with many well known chat-Bots too. Be it Slack or any other. One reason can be lack of **data**. Industries have websites full of text with product description or documents in a form of general reading. However, they lack when it comes to get relevant structural data for all their products and services for chat-Bot development. In addition to this, companies are evolving so fast that every now and then they get new products and services.Whereas, they lack to collect structured information about newly launched services to train the chatbot. Following are the research questions which is being addressed by present thesis.

- How can we tackle the data structurization for industry chatbot? i.e. how to convert unstructured data into structure or semi-structure one which can be useful for chatbot development. Is there any possibility to overcome from structured data crisis.

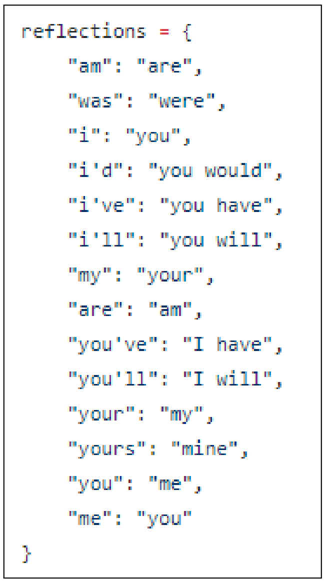
- How chatbot researchers can provide a chat-Bot development platform for unstructured corpus?
- What is the most significant effect of structure and unstructured data corpus on the performance of Chat-bots in industry domain?
- What are main development factors which occur while building a domain specific chat-Bot from unstructured corpus, and how can these commonalities be used to aid the industrial community in prevention of the problem?

2.2 Related Work / State-of-the-Art

Chat bots are around since 1966, when first Eliza bot was released. It was a rule based system which can find cue words or phrases in the input and reply as per programmed responses. For example if a query contains a phrase "friends" Eliza will ask you "tell me more about your friends"[1,4].This is how one can converse and it feels like you are talking to a human being. Eliza was not intended to deploy into industry domain rather the goal was to check if Eliza can fool human that they are talking to an individual. [cite-JosephWeizenbaum1966, Lorenz2017] Eliza works on structured data where we have reflection words and psychobabble sentences. Reflection words and Psychobabble sentences can be seen in following images.

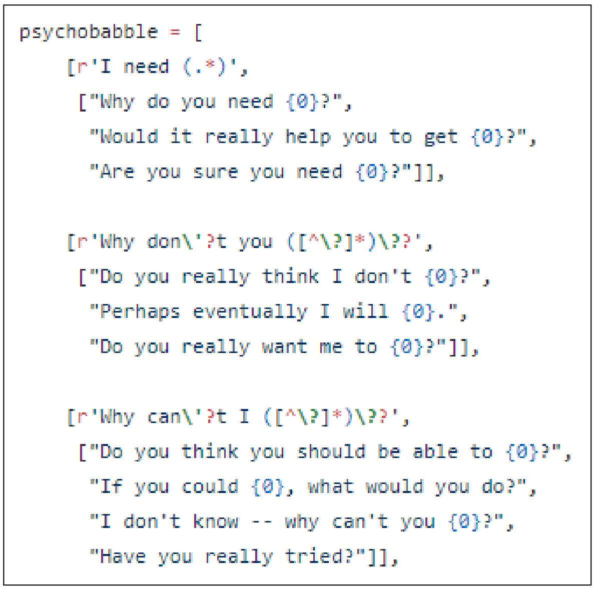
Looking at the data format one can see that Eliza repeat the given input and form meaningful responses so that an individual can feel that he/she is talking to a human being.

After Eliza, PARRY was introduced in Stanford University in 1995. It was also known an *Eliza with attitude*, developed by psychiatrist Kenneth Colby. Parry was tested on the eminent testing method Turing test. Where a group of experienced psychiatrists analyzed a combination of real patients and a computer program **Parry** using teleprinter. Another group of 33 psychiatrists were shown transcripts of the talk. The two groups were then asked to identify which of the "patients" were human and which were computer programs. The psychiatrists were able to identify only 48 percent of the time [2]. [citeColby1972] A.L.I.C.E. was then the most popular chabot based on Natural Language Processing. It was inspired by the first chatbot Eliza. Alice can converse with the user applying some heuristic pattern matching rules given a user's query. It was awarded as talking robot, humanoid and



```
reflections = {
    "am": "are",
    "was": "were",
    "i": "you",
    "i'd": "you would",
    "i've": "you have",
    "i'll": "you will",
    "my": "your",
    "are": "am",
    "you've": "I have",
    "you'll": "I will",
    "your": "my",
    "yours": "mine",
    "you": "me",
    "me": "you"
}
```

Figure 1: Reflection words for Eliza.



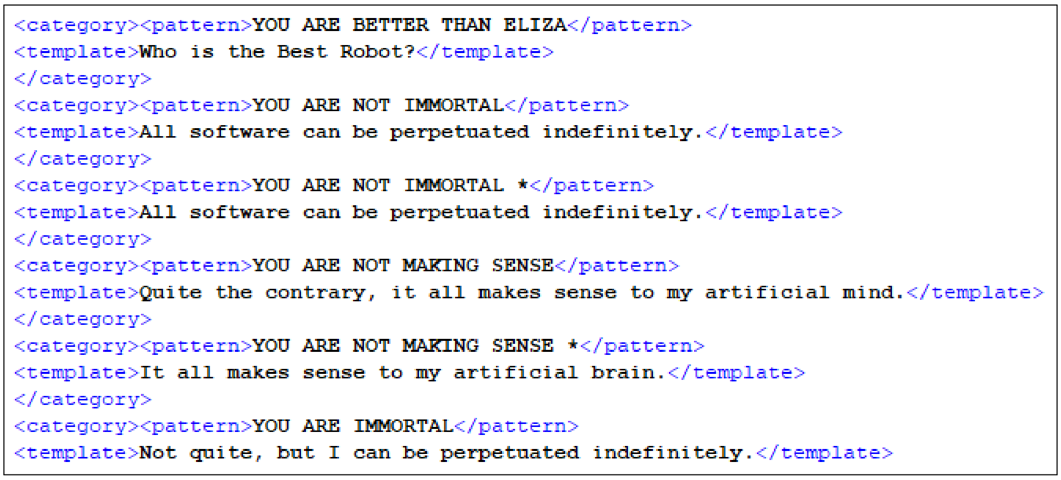
```
psychobabble = [
    [r'I need (.*)',
     ["Why do you need {0}?",
      "Would it really help you to get {0}?",
      "Are you sure you need {0}?"]],

    [r'Why don\'?t you ([^\?]*)\??',
     ["Do you really think I don't {0}?",
      "Perhaps eventually I will {0}.",
      "Do you really want me to {0}?"]],

    [r'Why can\'?t I ([^\?]*)\??',
     ["Do you think you should be able to {0}?",
      "If you could {0}, what would you do?",
      "I don't know -- why can't you {0}?",
      "Have you really tried?"]],
```

Figure 2: Psychobabble sentences for Eliza.

many more but unfortunately it failed the turing test [3,4].[citeRichard] The data for Alice is written in .aiml (Artificial Intelligence Mark-up Language) format which contains question followed by an answer. The format of the data which A.L.I.C.E.works on can be seen in figure. 3

Alice doesn't save the conversation history and it doesn't really understand what you said but respond you based on what she has in her brain.[citealexander2007] Right after the ALICE, industries started looking for the domain specific conversational system to talk to their customers, employees or with any other business. Haptik Inc. [6] (haptik2017) did a survey in 2017 on chat-Bot existence in the market. According to this report "currently exist over 40,000 chatbots across multiple platforms and that the market size of chatbots can grow from 700 million dollars in 2016 to 3 billion dollars in 2021". The report is from 2017, the numbers must have gone up by now. In 2016 Oracle Inc. survey says, that 80 percent of the 800 interviewed businesses were already using chatbots or planned to implement them into their businesses by 2020.

Ask Me Anything: Dynamic Memory Network for Natural Language Processing published in 2016. This research presents a new kind of network called *Dynamic Memory Network*for neural networks. It process input sequences and questions, form episodic memories, and generates relevant answers [5].



```
<category><pattern>YOU ARE BETTER THAN ELIZA</pattern>
<template>Who is the Best Robot?</template>
</category>
<category><pattern>YOU ARE NOT IMMORTAL</pattern>
<template>All software can be perpetuated indefinitely.</template>
</category>
<category><pattern>YOU ARE NOT IMMORTAL *</pattern>
<template>All software can be perpetuated indefinitely.</template>
</category>
<category><pattern>YOU ARE NOT MAKING SENSE</pattern>
<template>Quite the contrary, it all makes sense to my artificial mind.</template>
</category>
<category><pattern>YOU ARE NOT MAKING SENSE *</pattern>
<template>It all makes sense to my artificial brain.</template>
</category>
<category><pattern>YOU ARE IMMORTAL</pattern>
<template>Not quite, but I can be perpetuated indefinitely.</template>
```

Figure 3: Data Format for A.L.I.C.E.

Such as,
Sequence of input - Ankit went to hallway. Ankit picked-up the football. He traveled to the garden.
Questions - Where is football?
Answer - Garden

The most well known work has been done by Alibaba when they introduced AliMe chatbot for their costumer service. It helped Alibaba significantly during the peak selling season. AliMe is a *Sequence to Sequence* and *Rerank based* chatbot engine.[citeMinghui2017]. Later transfer learning and reinforcement learning were used to improve the performance of the bots in 2017-18 by Google and Alibaba researchers.[citeMinghui2018, YuWu2017, buck18] Currently, .json format of data is widely in use which can be seen **??** here.

In conclusion, almost all chatbots relied on structured data to train their model on. In some cases the bot were tested on the publicly available structured data to compare their methods with existing research.

2.3 Our Approach

"If knowledge to be extracted is expressed directly in the documents then *Information Extraction* alone can be used effectively for *Text Mining*. However, if the text information which we are aiming for is in unstructured text format rather than abstract knowledge, it might be helpful to use IE to convert the unstructured data into a structured database" [citeRaymond]. The present research will mine knowledge from unstructured text using information extraction and then giving extracted information a proper format. After getting the structured information, another challenge would be to label the intent of the users' question along with the entities attached to that particular intent. Mapping questions with the answers using TF-IDF approach might not work because of the ambiguity in the text corpora. Therefore, after extracting relevant information our attempt would be to label intent, entities and automate this labeling process rather than hand engineering [citeRaymond]. Our information extraction approach is slightly similar to the one of [citeRaymond]. When it comes to label queries with the intents and entities, there could be two approaches either using cloud API's or developing our own framework. Using cloud API's are the simplest and the easiest way to proceed.

2.4 Definitions

Chat-Bot - Chat-Bot(n.):
a computer program designed to simulate conversation with human users, especially over the Internet.(Oxford Dictionary)

AIML - Artificial Intelligence Markup Language
Bot - Chat-Bot or Conversational Agent
NLU - Natural Language Understanding
NLP - Natural Language Processing
NLG - Natural Language Generation
RL - Reinforcement Learning
TL - Transfer Learning
DB - Data Bases
KB - Knowledge Base
IR - Information Retrieval
DM - Data Mining
TM - Text Mining
IE - Information Extraction

3 Research Design

3.1 Data-set

Present research is going to use the unstructured data-set provided by COMPRA GmbH. It contains the detailed documentation of their product and services. They call them modules. We are going to choose a module and will clean the data and convert it into a format as such that we can train our chatbot on. After successful implementation we will add more modules.

3.2 Evaluation

For evaluation, we have adopted the way in which most famous chat-Bot **AliMe** was evaluated. COMPRA GmbH's analysts go through the answer of each testing question (two analysts for the experiment comparing with another public chat-Bot, and one for the other experiments), and mark them with three graded labels: "0" for unsuitable, "1" means that the answer is only suitable in certain contexts, "2" indicates that the answer is suitable. To determine whether an answer is suitable or not, we define five evaluation rules, namely "grammatically correct", "semantically related", "well-spoken language", "context independent" and "not overly generalized". An answer will be labeled as suitable only if it satisfies all the rules, neutral if it satisfies the first three and breaks either of the latter two, and unsuitable otherwise.

4 Organizational Stuff

4.1 Communication

Looking at the busy schedule of both the supervisor, I would request to have meeting once in every 2-3 week. I prefer 2 weeks though, so that we can proceed as quickly as possible. This will allow both of the supervisors to track my work and guide extensively. It's not necessary all the time to meet with all the COMPRA team but my personal meeting with Jonas will definitely help me a lot.

4.2 Presentation:

Generally professor Schmidt organize joint interim presentation for all the master thesis aspirant.The first idea talk is on 27th of November and next interim presentation date will be announce thereafter.

Table 1: Timetable

Time	**Activities**	**Results**
October 2018	Learning Latex, Literature research, Research questions	research proposal, related work, bibliography
November 2018	Data Pre-processing, Cleaning, extracting useful info	Semi-structured data
December 2018	First Implementation	Analysis of the result
January 2019	Further Implementation, improve existing implementation	Analysis of the result
Fabruary 2019	.....	....
March 2019	.....	....

4.3 Publication:

As we all know that we are attempting to tackle a task that has nearly no baseline (looking at the data format), this is the first draft of the thesis and beginning of the research. We all will get to know about the publication decision soon.

4.4 Provisional time-table

The table 1 will structure my whole work and research process. However this is not the final structure, since it still can change while working on the thesis and by acquiring new results and insights. Nevertheless it is a first attempt to generally structure my final thesis.

References

Laura Igual and Santi Seguí. Supervised learning. In *Introduction to Data Science*, pages 67–96. Springer, 2017.

Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics Reports*, 519(1):1–49, 2012.

Ken Peffers, Tuure Tuunanen, Charles E Gengler, Matti Rossi, Wendy Hui, Ville Virtanen, and Johanna Bragge. The design science research process: a model for producing and presenting information systems research. In *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006)*, pages 83–106. sn, 2006.

Coming Soon

I have already put all the references in bibliography section but not able to cite properly. Therefore, I am sorry here is no reference section. I'll soon add those.