

A Comparative Analysis of Generative Neural Attention-based Service Chatbot

Sinarwati Mohamad Suhaili¹, Naomie Salim², Mohamad Nazim Jambli³

Pre-University, Kota Samarahan, Sarawak, Malaysia¹

Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia^{1,2}

UTM Big Data Centre, Ibnu Sina Institute for Scientific and Industrial Research,

Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia²

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,

Kota Samarahan, Sarawak, Malaysia³

Abstract—Companies constantly rely on customer support to deliver pre-and post-sale services to their clients through websites, mobile devices or social media platforms such as Twitter. In assisting customers, companies employ virtual service agents (chatbots) to provide support via communication devices. The primary focus is to automate the generation of conversational chat between a computer and a human by constructing virtual service agents that can predict appropriate and automatic responses to customers' queries. This paper aims to present and implement a seq2seq-based learning task model based on encoder-decoder architectural solutions by training generative chatbots on customer support Twitter datasets. The model is based on deep Recurrent Neural Networks (RNNs) structures which are uni-directional and bi-directional encoder types of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). The RNNs are augmented with an attention layer to focus on important information between input and output sequences. Word level embedding such as Word2Vec, GloVe, and FastText are employed as input to the model. Incorporating the base architecture, a comparative analysis is applied where baseline models are compared with and without the use of attention as well as different types of input embedding for each experiment. Bilingual Evaluation Understudy (BLEU) was employed to evaluate the model's performance. Results revealed that while biLSTM performs better with GloVe, biGRU operates better with FastText. Thus, the finding significantly indicated that the attention-based, bi-directional RNNs (LSTM or GRU) model significantly outperformed baseline approaches in their BLEU score as a promising use in future works.

Keywords—Sequence-to-sequence; encoder-decoder; service chatbot; attention-based encoder-decoder; Recurrent Neural Network (RNN); Long Short-Term Memory (LSTM); Gated Recurrent Unit (GRU); word embedding

I. INTRODUCTION

Providing excellent customer service while engaging with their clients has become more pivotal than ever in today's digitally connected era. Companies engage with customers to assist them with pre-post sale items regularly upgraded due to technological advancements or the communication revolution. Over the years, face-to-face physical meetings and phone calls have been the two most dominant communication methods. Since the rise of the internet, various ways have evolved, from email to social media, installing the mobile application to fill out a form on a website, and eventually waiting for a follow-up. Recently, the increasing use of real-time messaging such as Twitter, Facebook Messenger, WhatsApp, Telegram,

Slack, etc., has led to a fundamental transition in how people would prefer to connect with businesses. While most of these communication channels have common characteristics, including online chat, which initially relies only on humans to conduct mutual communication, the baton now is passed to virtual agents or assistants called chatbots. Chatbots, the trendy platform led by virtual assistants, function as customer service representatives who negotiate conversations with clients to improve the user experience and services.

Chatbots are the subsequent major advancement in conversational services, which allow some business companies to communicate through messaging systems, like Twitter and Facebook Messenger, based on artificial intelligence and machine learning. Chatbots can be defined as computer programs living in messenger applications and providing specific services via emulating an interaction with a human through text messaging or a virtual voice [1] [2]. Owing to the overwhelming prevalence of chatbots as messaging is the most commonly used customer assistance medium; therefore, there is a need for the company to invest in a chatbot to support serving their customers' needs as applying in the context of service chatbots. Consequently, companies can strengthen employees' productivity to serve more customers with other services.

Chatbots' primary purpose is to facilitate the conversation between machines and humans in natural language conversation; as in the human viewpoint, these interactions should resemble humans as closely as feasible. Consequently, achieving this has become a fundamental task, with numerous researchers seeking the optimal way for having a chatbot to behave like a human. An effective chatbot should be able to comprehend the user's message, retrieve appropriate information according to the given statement and respond accordingly so that the user perceives the conversation as human-like.

The existing chatbots work just on pattern matching inputs and then finding a scripted answer corresponding to the information presented. The downside to this technique is that it cannot lead to a completely satisfying conversation due to the limitation of discourse within a specific domain with a clear goal. To handle the user's input utterances, Eliza, PARRY, and ALICE, to name a few were among the first chatbots to employ rudimentary parsing, pattern matching, or keyword retrieval approaches. These techniques require hand-written rules to generate responses. Due to the domain-specific