FlixMobility - Offline Home Task

0.1 Abstract

The call centres are moving towards digitalization to automate the process to enhance the reliable, user-friendly, and cost-efficient user service. The automation service helps the customer service employees in eliminating the repetitive tasks and reduces the work pressure by handling the complex situation, and also provides the revenue growth for the company.In this task, an idea has proposed to develop an system of agents to handle multilingual requests from the user.

The data is the crucial part to make the system effectively handle the tasks by training it respectively. The proposed architecture has the capability to adapt itself for multi-domain tasks such as restaurant reservation, hotel booking and travel information respectively or for machine translation tasks with multilingual options such as German to English, German to French, and German to Italian or else multilingual domain specific tasks. Despite the popularity of Transformer, and BERT models for Natural Language Processing which has the efficacy, the recurrent Neural Networks still plays a prominent role in handling Natural Language Processing tasks for automation because of its effective computation time. The proposed architecture is by integration of Progressive learning for learning incrementally with sequence-to-sequence architecture along with a Multi-head Self-Attention Mechanism. The detailed information has been provided in the forthcoming chapters.

Additionally, the challenges and the scalability of the architecture has discussed. The limitations of the proposed idea has been discussed later and a proposal for the challenges has been described.

Chapter 1

Introduction

Digitalization is paving a prominent way to automate the customer service sector, It Security, Industry 4.0 efficiently complying with the European Regulations. To automate the customer service industry, the business and research focused on the innovative development Mobil Dialogue Systems, Cognitive Systems, Personal Virtual Assistants, and Chatbots for applying in Automotive, Industrial Process Automation, Logistics and Block Chain technology. The emotional attributes of the call centre agents has a great influence on the customer end, thus mandate the high-demand in the delivery quality both from the customer and the industry side. Since the calls are either inbound or outbound, where the information sharing is happening between different geographical locations, difference in the gender, race, and the personality based on the nationality and the location. Hence, the call centre workforce undergoes lots of pressure in handling these complex situation [1].

Thus the development of Conversational agents has attained the business and the research focus for providing a cost-effective solution to enhance the revenue growth and thus enabling the reduction of pressure for call centre workforce. Though there exists a heavy competition in developing such a system, the demand from the customer end is high, and the business requirements gets challenging to compete for an effective solution. The system must be designed to be effective such that it is scalable, sustainable and adaptable.

1.1 Challenges in developing interactive agents

The customer requirements and demands are high requesting for a personalized approach in the customer service sector. Despite various expectations of each user, the system should be capable to provide the following deliverables [2].

- **Natural Interactions** The enquirer expects the system to behave naturally like a human, such that an interaction can be hold to achieve their goal from the system. To construct the conversation, the system must be capable to deliver a response according to the user experience with a clear grammatical structure.

- **Flexibility** The system should be designed in a way such that it can access multiple channels along with a adaptive techniques to handle multiple requests with multilingual options.

- **Responsiveness** A relevant response to the user's request is of much importance to make it contextually correct to proceed the conversation with the user. Also, an appropriate responsiveness signifies an individual attention for satisfying the users.

- **Information Clarity** The information delivered to the user should be clear, concise and precise in the system decided response. An apt quality, and quantity should be delivered.

1.2 Identified Problem in designing the Conversational agents

Despite the technological advancements in automation through Natural Language Processing techniques, there are limitations in interpreting the messages along with the associated intent and entities effectively, and an apt natural response throughout an entire conversation in achieving the goal. The data source is a significant factor in developing a effective system to handle the user's requests, whereas the technological strategies is a challenging task in building such an effective system. The challenges include,

- **Reliable Service** The customer and the service employee conversations involves confidential, and complex conversation, which creates high demands in the expectations and the requirements of the customer for a high-quality delivery. The system must be reliable at all the time, and adaptable for the multiple requests from the user.

- **Cooperative principles in handling the conversation** H.Paul Grice proposed four maxims cooperative principles, which states that to have a meaningful conversation Quantity, Quality, Relation, and Manner are more important [2]. The Quantity determines the information relative to the requests, quality defines the truthfulness in delivering the information as a response, Relation determines the relevant contextual information corresponding to the user's requests, and Manner represents the transparency in the data with proper ethics in delivering the response.

- **Responsiveness with responsibility** The system should respond with natural human-like dialogues to make the customers engage in the conversation. Hence, the system must request the user for additional information, provide relevant additional informations, and take the conversation in a desired direction. Also, the system must be capable to handle the generic information, handles improper requests, and clarify the misunderstandings, which either involves certain intrusive rules or else self-decisive techniques.

1.3 Objectives for the Task

The data has a crucial importance in designing the system. The resourceful data helps the system in providing natural responses while interacting with the users. The advanced automation techniques need to be employed in NLP techniques with scalability, adaptability and sustainability. The busniess logic integrations such as API calls and Management, business requirements and the customer needs provide an effective communication. The evaluation of the developed system by the live users for enhancement of the trained model.

The major Objectives are as follows,

- **Data development and Engineering** The natural human conversation helps in designing the system with maximum efficacy for reliable, and high quality service. The data should have the quality and quantity which meets the customer and business requirements.

- **Designing of the system** Advanced automation techniques must be employed in Natural Language Processing to interpret the user's requests, intent and the associated entities, selection of an apt system action and delivering a quality response with desired information, and ethics. The system could be integration of multiple agents for handling multiple tasks and requests with multilingual options complying with business and stakeholder requirements.

- **Integration of Business and Stakeholder Requirements** The system can either be integrated or separate which cannot handle two calls simultaneously, and has the capacity to handle four text-based tasks at the same time. The capacity is limited to three tasks if it attends at most a call along with two other text-based tasks. The priority of the language is German, followed by English, French, Italian, and others with the processing time of 10 minutes. The calls have always the priority to the text-based tasks.

- **Selection of tools to build the system** Analysis on different frameworks to select a suitable option to build conversational agents and dialog systems to cope up the complexities in designing such system.
- **Evaluation of the System** The system must have to be evaluated by live users and call centre agents before deploying it for the production. The Industry Acceptance Test must be performed through live users and it has to be approved by the team, industry and the customer for commissioning it.

Chapter 2

Proposed Idea and Implementation work

After analysing from the knowledge of the author, the proposed idea has the novelty, which has not been implemented yet. The idea behind this architecture is that it is an integration of progressive learning with sequence-to-sequence learning architecture for a progressive incremental learning, which has provided with multi-head self attention. By progressive incremental learning it can learn multi domain-specific data to serve user's requests from multiple trained domains. due to complexity in trainable parameters, maximum of three columns of neural network layers have used, and hence it is trainable for maximum three tasks or domain.

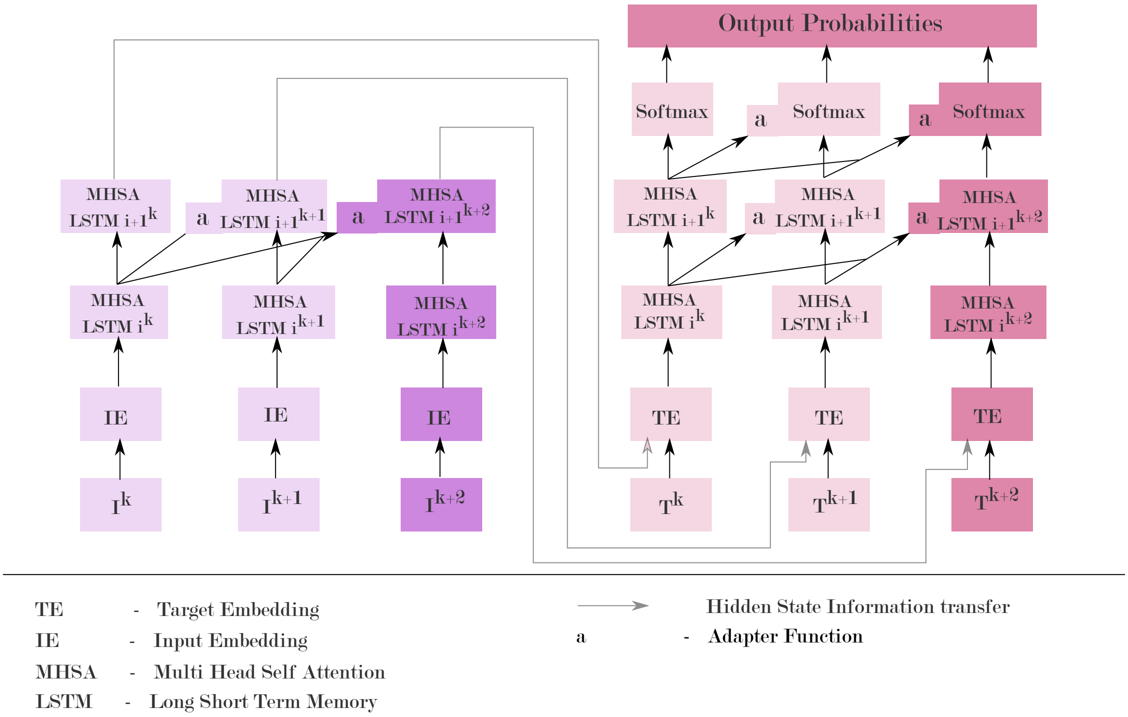

Figure 2.1: Progressive Multi-headed self-attentive Encoder Decoder Architecture

The advantage of this architecture is that the multiple agents have integrated as a single system to perform multiple tasks such as handling multilingual requests. After training the network, the saved model (separate for each channels) can be loaded in various channels to access it respectively. The column of feed-forward layers can be adjusted according to the convenience.

2.1 Gating Mechanism to customize the LSTM

$$\begin{multline} i_t = sigmoid\Bigg[U_i * hi_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{hi} * U_{vi}}{U_{hi} + U_{vi}})h_{t-1} * v_T + (\frac{W_{qi} * U_{vi}}{W_{qi} + U_{vi}})q_i * v_T \\ +(\frac{U_{ci} * U_{vi}}{U_{ci} + U_{vi}})ci_{t-1} * v_T + \sum_{k<j} sigmoid\Bigg[U_i^k * hi_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{hi}^k * U_{vi}^k}{U_{hi}^k + U_{vi}^k})hi_{t-1} * v_T \\ +(\frac{W_{qi}^k * U_{vi}^k}{W_{qi}^k + U_{vi}^k})q_i * v_T + (\frac{U_{ci}^k * U_{vi}^k}{U_{ci}^k + U_{vi}^k})ci_{t-1} * v_T \tag{2.1} \end{multline}$$

$$\begin{multline} f_t = sigmoid\Bigg[U_f * hf_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{hf} * U_{vf}}{U_{hf} + U_{vf}})h_{t-1} * v_T + (\frac{W_{qf} * U_{vf}}{W_{qf} + U_{vf}})q_f * v_T \\ +(\frac{U_{cf} * U_{vf}}{U_{cf} + U_{vf}})cf_{t-1} * v_T + \sum_{k<j} sigmoid\Bigg[U_f^k * hf_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{hf}^k * U_{vf}^k}{U_{hf}^k + U_{vf}^k})hf_{t-1} * v_T \\ +(\frac{W_{qf}^k * U_{vf}^k}{W_{qf}^k + U_{vf}^k})q_f * v_T + (\frac{U_{cf}^k * U_{vf}^k}{U_{cf}^k + U_{vf}^k})cf_{t-1} * v_T \tag{2.2} \end{multline}$$

$$\begin{multline} c_{new} = sigmoid\Bigg[U_c * hc_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{hc} * U_{vc}}{U_{hc} + U_{vc}})hc_{t-1} * v_T + (\frac{W_{qc} * U_{vc}}{W_{qc} + U_{vc}})q_c * v_T \\ +(\frac{U_{cc} * U_{vc}}{U_{cc} + U_{vc}})cc_{t-1} * v_T + \sum_{k<j} sigmoid\Bigg[U_c^k * hc_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{hc}^k * U_{vc}^k}{U_{hc}^k + U_{vc}^k})hc_{t-1} * v_T \\ +(\frac{W_{qc}^k * U_{vc}^k}{W_{qc}^k + U_{vc}^k})q_c * v_T + (\frac{U_{cc}^k * U_{vc}^k}{U_{cc}^k + U_{vc}^k})cc_{t-1} * v_T \tag{2.3} \end{multline}$$

$$\begin{multline} o_t = sigmoid\Bigg[U_o * ho_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{ho} * U_{vo}}{U_{ho} + U_{vo}})ho_{t-1} * v_T + (\frac{W_{qo} * U_{vo}}{W_{qo} + U_{vo}})q_o * v_T \\ +(\frac{U_{co} * U_{vo}}{U_{co} + U_{vo}})co_{t-1} * v_T + \sum_{k<j} sigmoid\Bigg[U_o^k * ho_{t-1}\Big\{softmax\Big[v^T * tanh\Big(\Big(\frac{U_{ho}^k * U_{vo}^k}{U_{ho}^k + U_{vo}^k})ho_{t-1} * v_T \\ +(\frac{W_{qo}^k * U_{vo}^k}{W_{qo}^k + U_{vo}^k})q_o * v_T + (\frac{U_{co}^k * U_{vo}^k}{U_{co}^k + U_{vo}^k})co_{t-1} * v_T \tag{2.4} \end{multline}$$

2.2 Advantages of the proposed architecture

Though Transformer and BERT architecture are efficient in its performance, it lacks in computational cost of time due to the stacked compilation of multi single head scaled dot product attention. Whereas in this proposed work, due to recurrent connections the computation complexity has reduced.

It has the adaptability for multiple tasks and multi domain setting thus providing the flexibility in accessing. It is also scalable to the factor of three domains to handle complex conversational tasks. Additionally, it is also sustainable due to its multi-tasks efficacy.

2.3 Challenges to face in the proposed architecture

Though it has the capability to learn from multi-domain, it does not have an option yet to switch the domain specific tasks in the end . Also, the complexity of the trainable parameters increases with the increase in the number of columns of feed-forward neural networks. Hence, it has been limited to three columns and two hidden layers to reduce the trainable parameters.

A system of parallel connections from all the domains simultaneously could overcome the challenge which might be faced during the evaluation.

2.4 Preferred tools

Tensorflow and keras offer wide range of Machine Learning and Deep learning layers to customize for the use case. It supports multiple python and backend libraries to support for the development of novel architectures. Rasa opensource might be a feasible option because of its flexibility, and adaptability in its framework to opt for various state-of-the-art NLP techniques for many use cases and success factors.