Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 5, 32-42 (2021)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Design and Implementation of Intelligent English Learning Chabot based on Transfer Learning Technology

Nuobei Shi, Qin Zeng, Raymond Shu Tak Lee*

Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, 519000, China

ARTICLE INFO

Article history:
Received: 21 April, 2021
Accepted: 27 August, 2021
Online: 10 September, 2021

Keywords:
NLP-based Chatbot
Transfer learning
OpenAI GPT-2
English Learning Chatbot
Artificial Intelligence

ABSTRACT

Chatbot operates task-oriented customer services in special and open domains at different mobile devices. Its related products such as knowledge base Question-Answer System also benefit daily activities. Chatbot functions generally include automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), natural language generation (NLG) and speech synthesis (SS). In this paper, we proposed a Transfer-based English Language learning chatbot with three learning system levels for real-world application, which integrate recognition service from Google and GPT-2 Open AI with dialogue tasks in NLU and NLG at a WeChat mini-program. From operational perspective, three levels for learning languages systematically were devised: phonetics, semantic and "free-style conversation" simulation in English. First level is to correct pronunciation in voice recognition and learning sentence syntactic. Second is a converse special-domain and the highest third level is a language chatbot communication as free-style conversation agent. From implementation perspective, the Language Learning agent integrates into a WeChat mini-program to devise three user interface levels and to fine-tune transfer learning GPT-2 as back-end language model to generate responses for users. With the combination of the two parts about operation and implementation, based on the Neural Network model of transfer learning technology, different users test the system with open-domain topic acquiring good communication experience and proved it ready to be the industrial application to be used. All of our source codes had uploaded to GitHub: https://github.com/p930203110/EnglishLanguageRobot.

1. Introduction

Chatbot operates similar to *virtual personal assistant (VPA)* and *question-answer (QA) system* in its development. In general, a practical chatbot consists of two categories, one has no *artificial intelligence (AI)* technologies i.e. rule-based and pattern recognition that collects tremendous high quality artificial corpus as database for *question-answer (QA)* matching [1]. Another has AI that uses current models and algorithms enabling chatbot to learn necessary articles before use. A typical example is speech recognition with text-to-speech interconversion to listen and converse at real-world human-machine interaction with database as machine brain or corpus with various meanings used as *natural language understanding (NLU), natural language generation (NLG), neural network (NN) models* with sufficient target language corpus training to equip the brain with thoughts and communication. This new human-machine interaction method provides convenience but it generates interior questions on how the chatbot can perform intelligently. It has been an ongoing research topic on user interface and expected functions by many internet organizations. However, back-end support require both AI and/or non-AI technologies to operate effectively. A literature review on human-like chatbot with more than one algorithm to select the best response that similar to Microsoft's Xiao Ice, a software assistant with emotions will be presented. A chatbot's performance is dependent on core components enriched with sufficient algorithms. Thus, a transfer-based technology with different language learning models would be proposed in this paper.

Turing test progress enticed researches on rule-based human-machine interaction system using pattern recognition in 1950s. Corpus as response matching outperformed Turing test showed

*Corresponding Author: Raymond Lee, Email: raymondshtlee@uic.edu.cn

that database can be enlarged by sufficient human dialogue directing that human syntactic simulation is not necessary as semantic matched natural human dialogue for response. Since then, when natural language is generated by trained machine with big data, the rule-based chatbot is updated to an *information-retrieval (IR)* or a *neural network (NN)* based system automatically and reduced manpower considerably. IR based system is mostly applied in a search engine optimizer with constructed knowledge base whereas NN system is densely data-driven to perform high-level prediction and classification. It can be used as a transfer learning method to pre-train and fine-tune the basic concepts of a chatbot. As shown on Figure 1, machine linguistic is depended on its data learned such as the basic level with words, phrases and sentences etc. For *natural language processing (NLP)*, an initial step is to transform natural language into a word vector followed by mathematical computation to illustrate the basic meanings equivalent to real-world. Thus, NN has the ability to simulate human brain and self- generate natural language as compared with other methods.

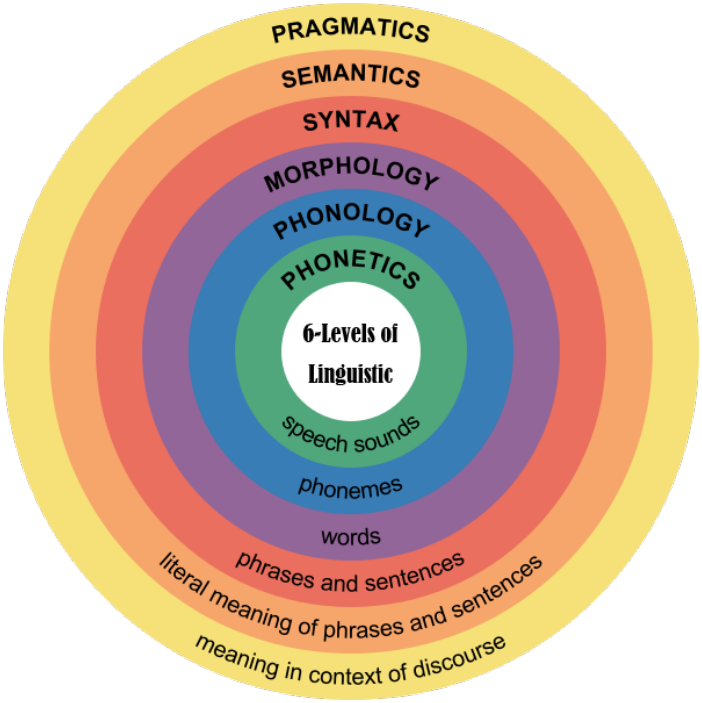

Figure 1: Levels of Linguistic in human languages [2]

There are numerous chatbot models equipped with different algorithms. Transfer learning is amongst performed better than various unsupervised learning sequence-to-sequence (seq2seq) models i.e. *recurrent neural network (RNN)* and its related bi-direction RNN and *long-short term memory (LSTM)* models. BERT from Google and GPT-2 from Open-AI are the most famous and pre-trained transformer models to generate natural language and classification. In this paper, GPT-2 is used as English language learning back-end model to converse daily dialogues because of GPT-2 inherent priority of natural language generation. For pre-trained transfer learning language model, a small dataset is used to fine-tune the transformer model to outperform traditional counterparts. Also, a self-attention mechanism is included. It is the main theme of transfer learning to find the relationship between sequences. Further, a language mini-program based on GPT-2 with a fine-tuned model will be included to provide fluent and syntactic responses. Lastly, a mini-program equipped with a fine-tuned GPT-2 language model and speech recognition service from Google to deliver a three systematical English learning level functions at WeChat for implementation.

2. Literature Review

2.1. Overview of industrial English Learning apps

Duolingo is an English language learning application with growing popularity since coronovirus. It is accredited by universities in UK, US and Australia as English language proficiency qualifications similar to IELTS and TOEFL. It contains spoken tests' pronunciation and oral presentation on text questions or images understanding with scores given depend on voice recognition. Test contents are preinstalled without human participation and gradings are given by statistic algorithm that are different from IELTS by physical examiner. It covered more than 30 mainstream languages transmitted in English and more than 6 languages relayed to Chinese language. Prior course commencement, the application would match users' demands to set step by step learning schedules and provided tests to detect grammar and pronunciation proficiciency. Level detection is an important component to improve language skills because it depended on applications or chatbots functions and users' willingness to select appropriate learning levels at different courses. [3] It shows that the online English learning or human-machine interaction English test will be the mainstream in these days. However, the technologies it used are not clear or transparent for users, the statistical algorithm or non-AI technologies also cannot substitute the official recognized IELTS which is communicate with real human or the machine recoreds judge by human. On the contrary, our model are natural based on AI technologies and growing with the interaction between human and machine.

The model we proposed is based on IELTS Part ABC speaking module examination syllabus covering pronounciation, topic discussion and free chat. They are named as Level 1-3 where the highest level 3 is a human-machine interaction dialogue. As responses from reverse party generated by IELTS are in random so the same format is followed where responses are generated by a back-end transfer learning model similar to a proficient English Tutor. Corpus of big data used for transfer learning model training covers topic knowledge discussion.

Other reference is based on a different English language learning application called LiuliShuo. It contains 7 levels to improve verbal skills. Level 4 refers to proficiency equivalent to College English Test (CET)-4 or 6, where the highest level 7 is equivalent to IELTS 7.5 and TOEFL 105. LiulShuo settings, compared with Duolingo in user interface and functions design are fitted to the model design. However, they are sufficient for grades evaluation but are lack of AI advanced intelligent technologies for response according to user's learning progress.

Thus, the design of the proposed model began with daliy dialogue for model training to improve its learning ability, then a database was added to track every users' learning history and levels improvement. For this segment, daily chitchat data in natural langauge are transferred into json format to improve machine learning. Those unstructed raw data or called heterogeneous data

required cleaning and sorting automatically in accordance with AI ecosystem.

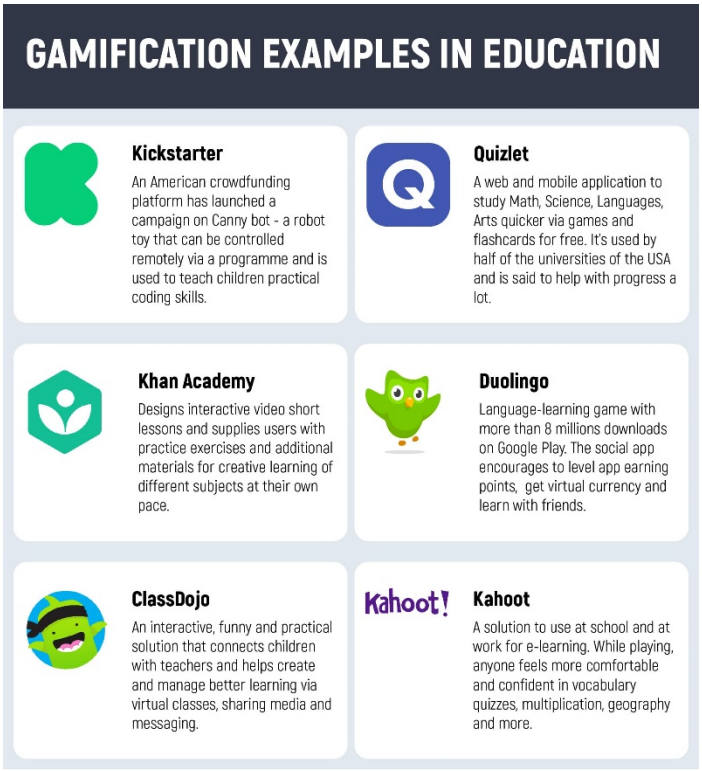

Figure 2: Gamification examples in Education [4]

Figure 2 shows a set of gamification examples in education. It is a game module to calculate grade levels to motivate users compete within app for improvement.

2.2. Academic research about Chatbot.

2.2.1. AliMeChat: A Sequence to Sequence and Rerank based Chatbot Engine (rule-based IR+seq2seq) [5]

Recurrent neural network (RNN) is capable to generate responses from end to end leading to sequence to sequence (seq2seq) model becomes chatbot mainstream generation prior 2018. Since then, NLP tasks were greatly improved using transfer learning model.

TaoBao App uses AliMe to substitute online human customer service for most of the predicted questions. Statistical data showed that 5% questions were chats from a commercial-domain Question-Answer system to an open-domain chatbot. This process resembled to English language learning agent similar to level 2 scenario dialogue to level 3 open-domain free style conversation of this paper. AliMe integrated a rule-based IR and pre-trained seq2seq application model to real-world industry and performed better than both IR and public chatbot. However, its pre-trained seq2seq model is used twice for response generation to re-rank a set of answer from matching input questions with knowledge base paired dialogues. It suggested that free style format can be based on usual conversation scenario exist in daily activities.

The proposed model has two similar AliMe settings in generation for information retrieval but in different formats. IR-based model uses natural language word to match knowledge base, the generation seq2seq model and model evaluation to re-rank output response in generation are embedded words as vector. IR-based dataset is approximately 9,164,834 natural questions and answers paired dialogues from users and staff collected from commercial domain. Researchers use inverted index based on search engine concept to match these conversation with input sentence containing identical words and used BM25 algorithm to measure input sentence similarity to select questions and respond to input questions. Traditional chatbot back-end also avoided commonsense questions that cannot be responded by chatbot and generation model used seq2seq model.

AliMe also selected GRU, a type of RNN units to reduce computation time for response generation. Further, an optimizer SoftMax is used in calculate to time control sample words set covering 512 random words. Beam Search in decoder assisted to find the highest probability in Conditional Probability and obtain the optimized response sentence within parameters. Results performance showed that *IR+generation+rerank* shown on Figure 3, by seq2seq model and mean probability scoring function evaluation approach achieved the highest score compared with other separated methods.

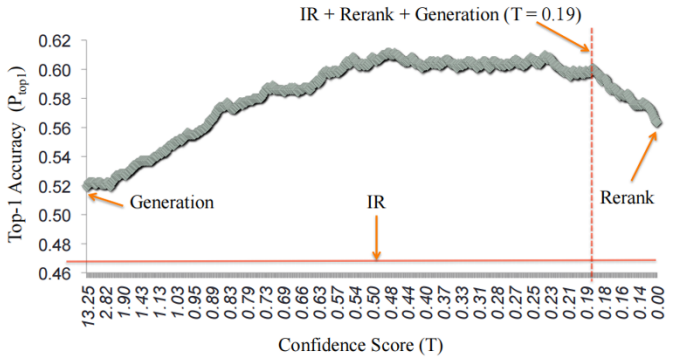

Figure 3: Top-1 accuracy of candidate approaches [5]

Due to the characteristic of NN and chatbot usage or a QA-based system, special-domain system in contents apart from the task-oriented or goal-oriented system for special usage than open-domain system applications are essential for model design. Microsoft chatbot is open-domain with emotions and AliMe specialized in business as Taobao customer service restricted in business corpus. Also, grounded knowledge contents becomes another important trend for chatbot developments[6]. In summary, rule-based dialogue with seq2seq RNN model are used frequently than others because most daily conversation cannot be generated by RNN model. Since attention mechanism proposed in 2017 in Attention is all you need [7], language model in natural language processing changed to *seq2seq+attention* and self-attention in transfer performed better than seq2seq.

2.2.2. The Design and Implementation of Xiao Ice, an Empathetic Social Chatbot（IR+seq2seq+KG）[8]

Microsoft released an empathetic social chatbot based on two segments at various platforms. One is intelligence quotient (IQ) that used for thinking and answer human questions. Another is emotional quotient (EQ) to understand and analysis the emotional meaning of natural language to provide 24 hours user services, research team use three back-end to equip the machine with human-like brain to realize these two segments. First is the traditional IR with corpus for matching. Second is a RNN of sequence to sequence model to generate response and the third is a knowledge graph for entities extraction to extract related entities

and sorting for responses. After that, an optimizer would rank potential responses and select the most suitable response to users.

Each Microsoft operation system has a unique Cortana for its computer owner [8]. Xiao Ice has different Cortana as a personal assistant resembled to Siri. It preferred to chat with IQ and EQ system design similar to human interaction. It is by far the most dialogue-oriented AI robot with emotions to provide 24 hours interaction. It has over 660 million users since its inception with *conversation-turns per session* (CPS) reached 23 which is higher than other chatbots. It is also an open-domain chatbot aimed to optimize content responses with emotions to understand users' thoughts. It has expanded to more than 5 countries with different names. Users willing to use not only depended on IQ but also dialogue contents provided a user-friendly first perception. Thus, IQ part would also be used in the proposed model.

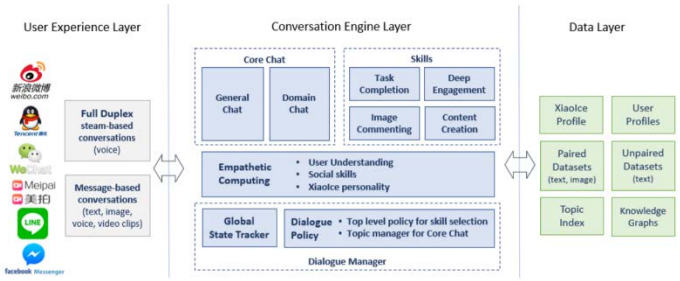

Figure 4: XiaoIce system architecture [8]

In Figure 4, for users implementation, Xiao Ice exists in 11 mainstream platforms including WeChat and Tencent QQ and at mini blog, Weibo and Facebook. It has the highest recognition function that include text, image voice and message-based conversation. Its conversation engine layer was developed after four years and its Core Chat can distinguish and change topics automatically. Thus, the proposed model is set with a second level at topic dialogues for special domain scenario and a third level for free style conversation to improve users' dialogue proficiency. A dialogue manager is to define dialogue management policy so that it can monitor the state of dialogue consisting Global State Tracker and Dialogue Policy to select a state action. Global State Tracker is a vector consisting of Xiao Ice responses and analyses text strings to extract entities and empathy. Dialogue policy is designed for long-term users to improve interactive engagement optimization. Initially, topic manager is the first to greet users and awake system with natural language understanding (NLU). It will then obtain users' responses' to adjust the "Core Chat" into open-domain or special scenario with users' interests. If system always repeats or response information is bland and within three words, the topic will be changed accordingly. Secondly, is the skill selection. Once user input is a particular format, the skills would activate to process different input. Its images can be classified into different task-oriented scenario e.g. if the image is food, a restaurant will be displayed. Lastly, a personal assistant would provide information such as weather, space availability and reservation for task completion.

In other words, it contained certain knowledge graphs in data layer where dataset is originated from popular forum such as Instagram in English or douban.com in Chinese. These datasets listed as several topics with relative small knowledge base as candidate answer and are sufficient for questions. When a new topic appears, system would obey the policy to refresh its knowledge base by machine-learning. However, not all new entity or topic will be collected. It is similar to a ranking list unless the entity is contextual relevance or the topic has higher popularity or news latest update and users' personal interests will be considered.

The core of Xiao Ice is language model, using the RNN shown in Figure 5 to create the open and special domain knowledge.

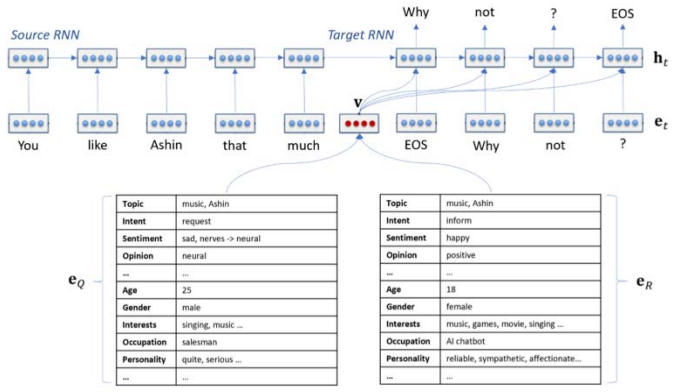

Figure 5: RNN-based neural response generator [8]

[Human] How old are you?
[S2S-Bot] 16 and you?
[Human] What's your age?
[S2S-Bot] 18.

[Human] How old are you?
[Xiaoice] I am 18, of course.
[Human] You age?
[Xiaolce] 18. Why?
[Human] Were you 18 last year?
[Xiaoice] I made a wish to stay 18 forever. Oh, my wish has come true.

Figure 6: Examples of inconsistent responses generated using a seq2seq model [8]

Figure 6 shows a chatbot with emotional consciousness. It enhanced interaction to improve system EQ performance.

Xiao Ice, in general, has retrieval and generated-based back-end models. The former is a mature search engine used for decades. Its knowledge graph is a database for graph matching and entities extraction where the latter is trained by big data with related dialogue based on entities relations to supply contents and explain output sentences from NN. These two directions using neural model to generate responses and rank sentences to match relations and optimization. Once entity recognition is linked with neural generator conversation and dynamic knowledge graph and the knowledge-based will be refreshed. If user's feedbacks match with English syntax, the knowledge-based would always be collected together with users' interests. In addition, a rule-based chatbot with *artificial intelligence markup language (AIML)* as backup will be included to avoid senseless or repeat answers.

Thus, cognition, reinforcement learning of history dialogue[9] and empathetic computing are essential segments in the proposed model for competitiveness.

2.2.3. Attention Is All You Need [7]

Although chatbots mentioned above are acceptable for usage, its training back-end testing about sequence to sequence model showed good performance for text generation. Considering neural network relied on big data, data-driven model for better performance. Thus, transfer learning model shown on Figure 7 with magnitude training and fine-tune data has been the architecture mainstream. With the recognition of attention mechanism, the updated self-attention created the priority of text generation to be proved by BERT and GPT and GPT has natural advantage for text generation within sequence.

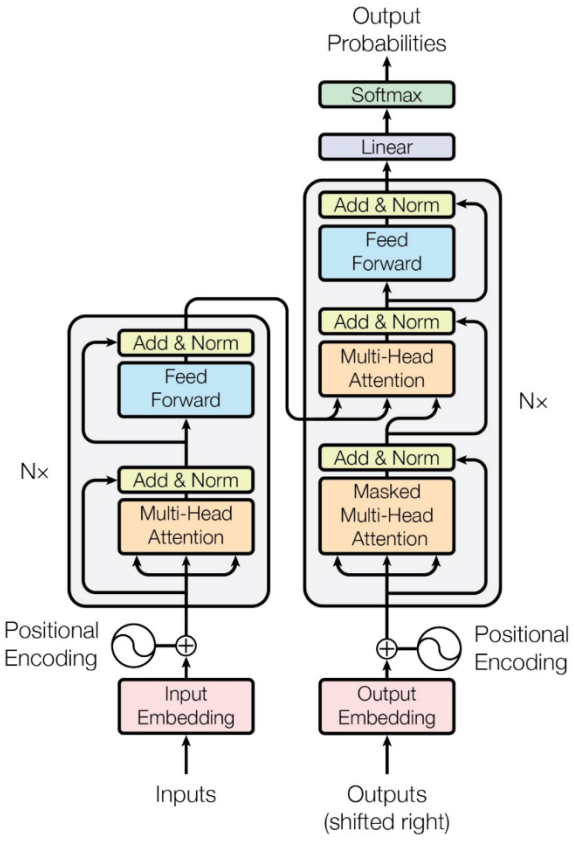

Figure 7: Transfer learning-model architecture [7].

The proposed model used as back-end is *Transfer Transfo*[1], as a daily dialogue system with common sense which integrate transfer learning training architecture and high-quality dataset. By fine-tuned persona-chat, the original utterance of GPT-2 with text generation based on context changed to format dialogue with fine-tune content dataset and English skills with original magnitude GPT-2 model. Also, the fine-tuned dataset also added the contents of English Tutor with users' personality and background resemble to native IELTS examiner.

Attention is human preference as a weight added to other words computation. It means that the encoding related to context, input and output. Transformer use self-attention that is search relations within sequence finding dependency among words in one sentence.

BERT is the first system proposed in transfer learning and fine-tuning. It has widely used in especially encoder part of pre-trained language model. However, it cannot predict decoder outputs. Open-AI GPT-2 has better application for generation with corpus and decoder limitation. The most famous performance of GPT-2 is text generation, since it releases larger dataset, more users are willing to make text prediction and dialogue. With the same transfer learning mechanism, GPT-2 not only can build dictionary use tokenizer like T, also play a role of language teacher with pre-trained grammar in natural language processing. Also, the GPT-2 original training source is 40G text from Internet, it has better adaption for text generation.

When the first edition of GPT-2 open source, researchers would like to test its performance with larger data. The 40G text in open-domain enabled GPT-2 equip with sufficient common sense similar to AIML(Artificial Intelligence Markup Language) which consist of human daily dialogues, but with more contents like grounded knowledge of Microsoft chatbot. For researchers, fine-tuning is the process to make GPT-2 transfer as own language model with special tasks. Text generation, text classification, automatic summarization, music modeling even writing code, all are functions of fine-tuned GPT-2 by users.

Thomas Wolf and his partners applied GPT-2 into dialogue generations called *Transfer Transfo* [1] to develop conversational system. Besides rule-based and seq2seq dialogue system, Transfo is another data-driven dialogue agent based on GPT-2, it combines transfer learning method to training and the transfer learning model with abundant contents with high-capacity, which shows big progress compared with end to end like seq2seq and information retrieval models. For traditional language models, neural network cannot track dialogue history within the model, even it learns the conversation with users. It is only big data to learn which cannot be tracked like rule-based system with database to store history dialogues. For training and generation algorithm, Transfo has priority to improve in relevance, coherence also grammar and fluency for output. For the model, multi-layers with 12 decoders and masked self-attention heads that token only range from left context. The large transfer learning model used in several practical task pays more attention to decoder part for the down-stream NLP tasks [10]. Pre-trained is an important part for down-stream performance to avoid discrete sentence to construct systematical learning in grammar and continuous content, document level corpus is better than sentence level. So, an Open AI release different size corpus for researchers to select suitable pre-trained and fine-tuning model afterwards. Fine-tuning decides the down-stream task text format in this paper. Thomas Wolf use persona-chat dataset as fine-tuning data to train input and output utterance change from long-text to dialogue format. Persona-chat in real-world helped to shape the speakers' backgrounds to make system further define topics and better understand users' input sentences .

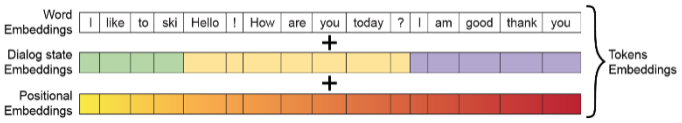

Figure 8: TranferTransfo's input representation [1]

In Figure 8, tokens embedding in one sequence is consist of 3 embedding of word, dialogue state and positional as shown in Fig 8. Rather than simple end to end dialogue with only one sentence, it collects usually 4-6 sentence and 3-5 of history dialogue belongs to same user in one sequence for training helping system to judge , state and position representation of every token to predict the next sentence in real application.

A dialogue system based on GPT-2 is included in the proposed model. In [11], the author made dialogue task more specific to goal-oriented personal assistant to help users complete special tasks. For task-oriented training, it has higher quality for corpus to fine-tune model and track conversation state of various situations. Not only it can interact with external system rather than pipelined system but also explain why the system generate such response.

End-to-end neural architecture for dialogue systems built on the top of GPT-2 with dialogue format. So, the dataset for fine-tuning should be dialogue as shown in Figure 9, which include external database of restaurants, hotels and others scenarios traveling in UK. The dialogues link to database contents to predict the system dialogue state and system action. Even more, following

the traditional end to end pipeline plus the database to be more explainable in output sentence that show why the system provides direction or operation advice as responds to users. Although it is a simple gradient descent, GPT-2 pre-trained model already equipped competitive dialogue as response.

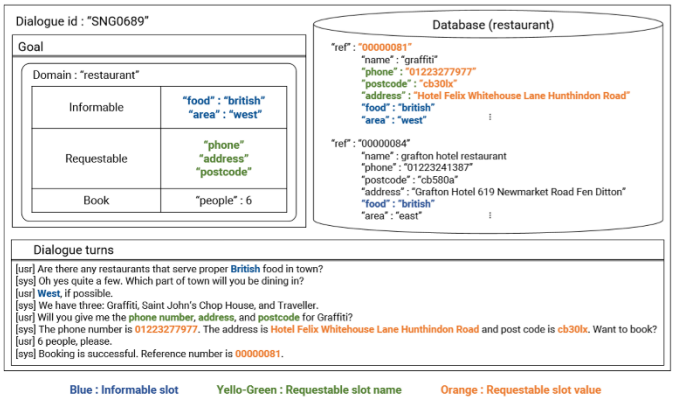

Figure 9: A single-domain example in MultiWOZ dataset [11]

Figure 10 shows the architecture of model similar to the model of grounded knowledge by Microsoft. Because the external system also a supplement for normal case which is a sequence to sequence model to generate substantial knowledge. When transfer learning-based system requires contents to engage specific task, knowledge base to fill the slot is another method to accomplish same object.

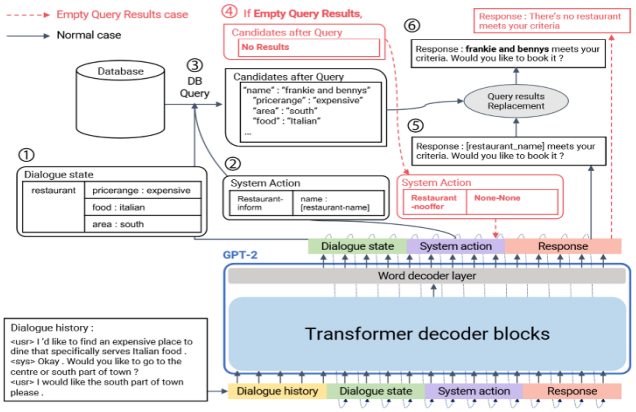

Figure 10: Architecture of end-to-end neural dialogue model[11]

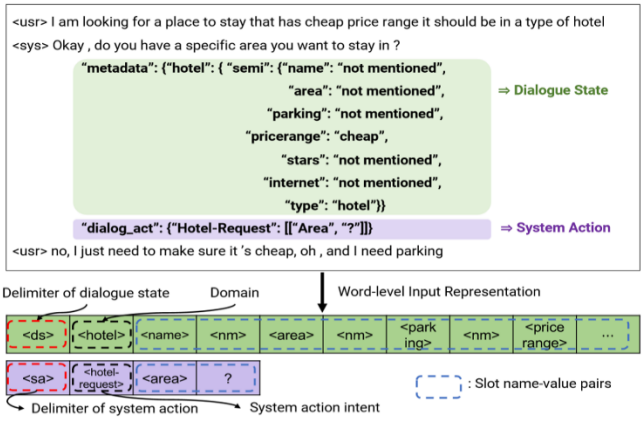

Figure 11: MultiWOZ dataset for test [11]

Dialogue state and system action, shown on Figure 11, is fundamental standard for the procedure of architecture. With personal assistant setting, this model also requires decoding strategy to optimize output sentence conditional on probability distribution. In this paper, top-k and top-p sampling is better than general beam search which are not suitable for high-entropy NLG and greedy decoding get highest probability over all probability such as for machine translation. To some extent, even though the database requires query to get real-world information tokens, there are not all sentence need external system to fill the slot, GPT-2 model will consider the situation to handle empty query for common sense responses and completion [12].

So far, progress in dialogue system in recent years due to the development in language model of natural language processing. In fact, most of techniques are for machine translation, sequence to sequence model and attention mechanism also the ones for machine translation. Until GPT-2, which natural has ability to generate text when the model is tested in below section. So, GPT-2 with transfer learning changeability for downstream natural language tasks with fine-tuning. Thus, we will use GPT-2 will be used as language model in the proposed model to obtain better response than seq2seq models.

After investigation and test of industrial products and some research projects about chatbot, with the development of AI technology, it was decided to use frameworks and transfer learning and fine-tune chatbot for English language learning. Chatbot in the proposed model will be a response generation chatbot with own common sense - the highest level is *free-style conversation agent* of a language system superior than other popular applications.

3. Methodology

In view of system framework, the design is firstly complied with the chatbot general flow at Figure 13, to decide whether it is functional. A top layer would be added followed by the application layer for systematically learning English which include three levels at User Interface designing scheme. As shown on Figure 12, the lower layer of architecture is the back-end model and the usage of language model using GPT-2 as language model responsible for the natural language understanding and natural language generation.

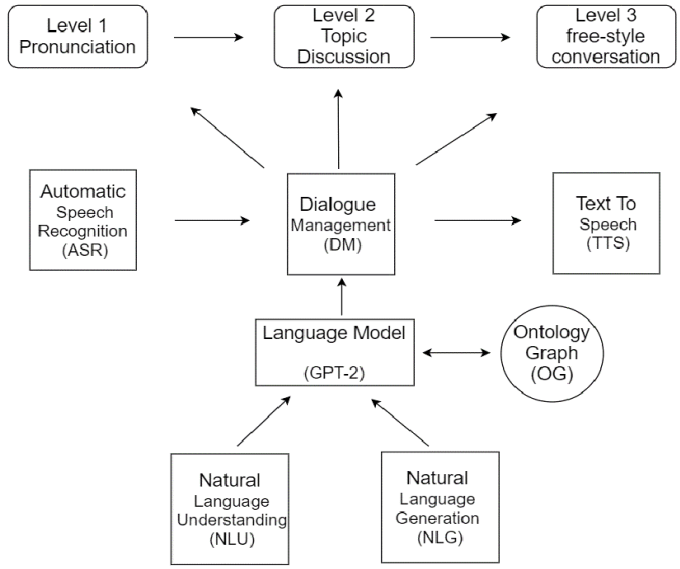

Figure 12: English learning chatbot architecture.

3.1. Chatbot system flow

In general, chatbot requires five parts to make communication with users and responses. The first step is voice recognition then change to text version natural language via ASR, the Automatic Speech Recognition is the same from text to voice in last step. For the central part, chatbot should handle natural language understanding and natural language generation using language model, it requires plenty of dataset as corpus to pre-train and fine-tune back-end model. About the optimization for daily dialogue and language model, with MySql database support the track of history dialogue distributed the user individually, the chatbot will become more intelligent and easier to modify with the long-term users.

Figure 13: System Flow of chatbot [13]

With the extension of system flow, in our architecture, the application layer are all English learning levels, which detail functions shown at implementation parts with mini-program User Interface and Develop Tool.

3.2. Open AI GPT-2 language model

Before we talk about Open-AI GPT-2, we should illustrate what is *language model (LM)*. For example, input editor is a language model which can predict the next word by language habits for behavior. From that, we can consider GPT-2 as a connection function of input editor. So far, GPT-2 already enlarge it language model with different size covering different needs of industrial NLP tasks.

Compared with the sequence to sequence we mentioned at section 2, the *transfer learning model* could provide better response that are more related to dialogue questions. The pre-trained dataset of Open AI GPT-2 are 40G text, it is no doubt that Open AI GPT-2 perform better than sequence to sequence model as it is data-driven neural network. The encoder part of GPT-2 will helps in the generation of natural language. Figure 14 shows the different size open source GPT-2 model for download and training.

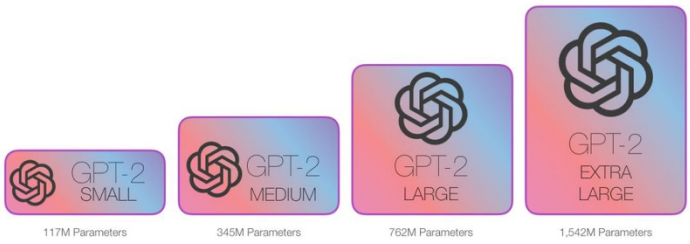

Figure 14: Size of Open AI GPT-2 pre-trained model [14]

Original GPT-2 always predicts the next token based on the information only at left hand of whole sentence. In summary, on probability condition, it uses history information with sufficient vocabulary to predict next token. After one prediction, it will provides a list for select, that is algorithm of top-k. In BERT, GPT-2 is also an architecture of transfer learning. Transfer learning consists of encoder and decoder, BERT focuses on the encoder and GPT-2 deletes encoder but keep decoder. Without encoder, decoder layers will be accumulated and magnified data. The idea of self-regression used by GPT-2 resemble RNN where it puts the token generated at the back of sequence, which the sequence will be the new input for the condition of next token generated.

In Figure 15, the encoders and decoders have 6 encoder and decoder. It removed traditional RNN and the pre-trained model fine-tuned by researchers will become mainstream in the future.

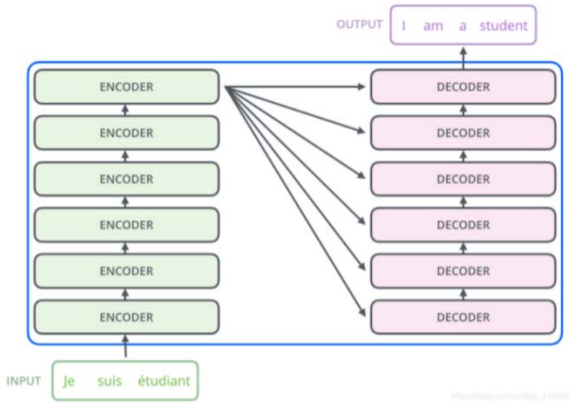

Figure 15: 6 layers encoder and decoder [14]

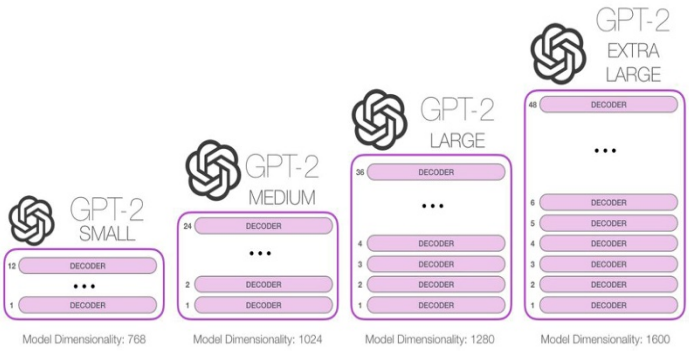

Figure 16: Decoder layers of Open AI GPT-2 [14]

The transfer learning change from BERT, shown on Figure 17, to GPT-2, the self-attention changes to masked self-attention shown on Figure 18. In Figure 16, due to GPT-2 have no encoder, input sentence as reminder is incomplete. The empty part of BERT covered with <eos> and other characters but GPT-2 with self-attention masked. So, it is masked self-attention. The impact for computation is only calculated the left words of token. The illustration of two different self-attention mechanism is shown in Figure 19.

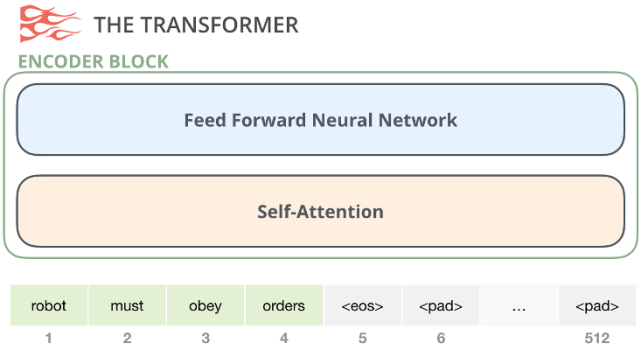

Figure 17: Encoder block of Google BERT [14]

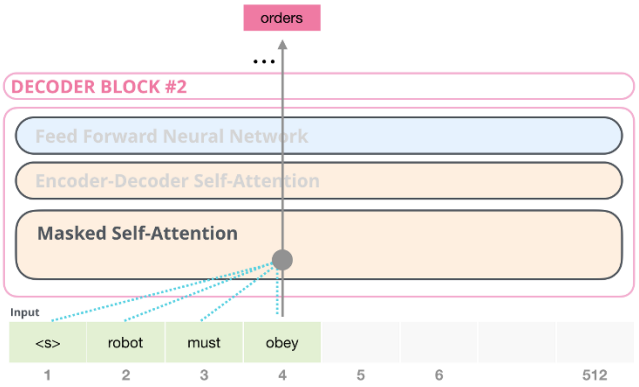

Figure 18: Decoder block of Open AI GPT-2 [14]

The masked self-attention GPT-2 used shows different with self-attention is that the masked one only know the left part of sentence, which include all information on left to predict next word to form the complete sentence even with articles. In addition, the token position encoding and embedding also considered into computation. From the start point of masked self-attention to the neural network to calculate the next word, the process will retain all the weight for the score of final values.

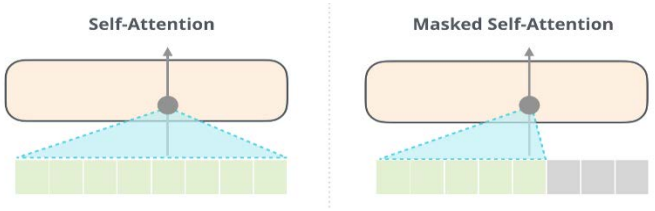

Figure 19: self-attention mechanism[14]

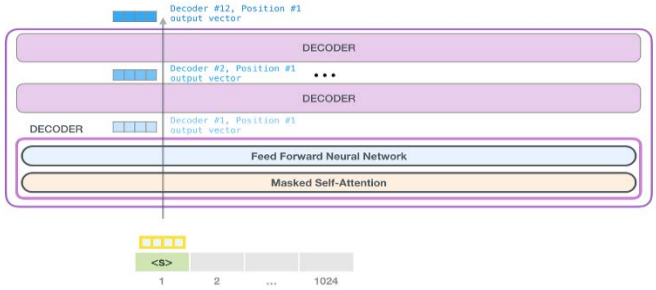

Figure 20: stack in Open AI GPT-2[14]

Figure 20 and 21 shows the process of prediction words. Sending a word to the first transfer learning block means looking up its embedding and adding up the positional encoding vector for position.

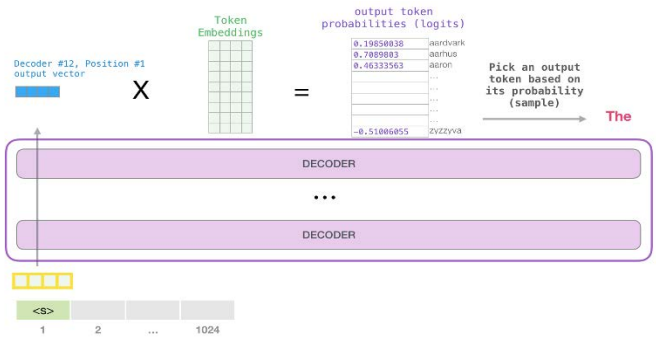

Figure 21: Open AI GPT-2 how to predict word in computation [14]

Initially, end-to-end and transfer learning are designed for machine translation with promising results. Since translation and dialogue are all sequence, more and more users used the same method to dialogue system. For research, only fine-tune with magnitude pre-trained model can obtain ideal language model for special tasks such as text classification, dialogue system etc.

4. System Implementation and Experimental Test

For the implementation of three levels English language learning, for users to access the system anytime, the platform of mini-program at WeChat for the front-end as User Interface to integrate our fine-tuned GPT-2 model based on the Google cloud server will be used in the proposed model. To realize the top layer applications of global architecture, Google speech recognition server is used as the recognizer of pronunciation due to Google has mature technology in speech recognition as compared with others. Level 1 includes the sentence distance comparison of recognized voice and sample sentence. Levels 2 and 3 use generate based language model to provide response to user. Level 2 is topic discussion from daily scenario with 8 topics is the representative of practical English. The highest level is called *free-style conversation* without any limitation for themes. In addition, the Open AI GPT-2 model also could process response to avoid meaningless response with common sense of English syntactic and semantic.

In Figure 22, our *user interface (UI)*, English learning agent has three levels for spoken English training. In this section, system usage and the different interaction performance levels with human users will be presented in this section.

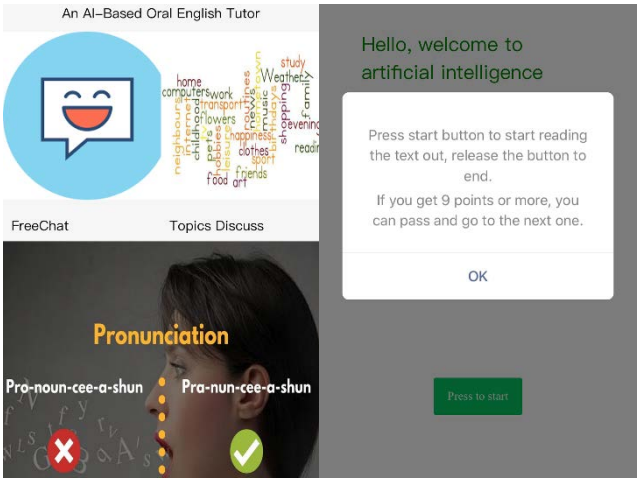

Figure 22: UI of English Learning Chatbot and Level 1 Pronunciation.

Pronunciation is the most important in English learning. This module is designed to practice. Once chatbot can recognize your voice to text and match the text read by user meaning that pronunciation is correct and can be recognized by a native speaker.

Firstly, the rule of pronunciation is the evaluation score shown on system more than 9 where the score comes from a native speaker's recognition set by Google speech recognition server. If user pass the pronunciation meaning that the words from users can be clearly recognized by voice recognition and match with the sample sentence. With the computation of voice recognized words and preinstalled sample sentence by *sequence matcher algorithm*, the essence of distance between two texts can be also calculated to distinguish the difference between user and the native speaker.

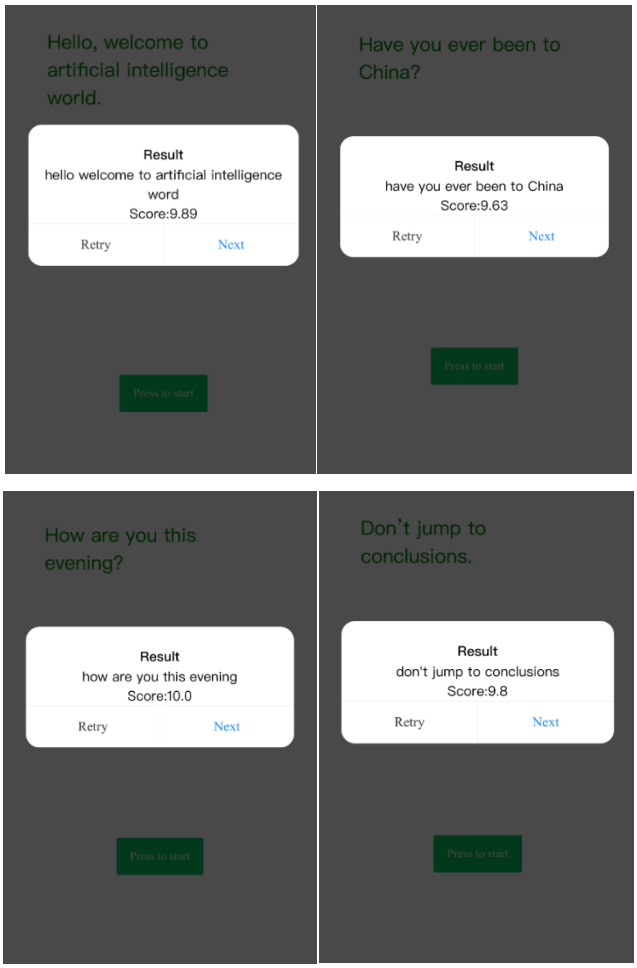

Figure 23: Test of Level 1 Pronunciation

During test in Figure 23, the first level only set 4 sample sentences as demonstration. After several rounds of test, the system gives higher score when user's voice is slow enough to distinguish every phonetic. Also, the testing or practice environment should be silent to avoid some acute noises. In addition, when different words with approximate phonetics such as 'word' and 'world', it will reduce the score pronunciation test in level 1, but the system may not transform voice to text accurately at level 2s and 3 which will influence the feedback from chatbot.

Second part is the process of a transfer learning model based on an Open-AI GPT-2 architecture so that the pre-trained model is able to generate high-quality text. The fine-tuned model used is *Transfer Transfo* where fine-tuned dataset is an open-domain daily dialogue. During the test, the UI settings are designed for dialogue scenario restricted in daily dialogue such as weather, travel, movies and other daily related activities.

For data training, the character always talks about movie, however, the range of movie contents are limited in US, with the influence of voice recognition. It is difficult to converse about the of movies contents. The model proposed has GPT-2t characteristic which according to the several reminders to write an article. Even the dialogue format limits the sentence length, it also follows the words from users to provide response and take no account of occasionally wrong input and fantastic output. Thus, the whole dialogue is worthy for further research. As mentioned, chatbot only focus on chatter. It has little knowledge base to generate meaningful sentences. So, the conversation contents cannot be limited within the topic as shown on Figure 24.

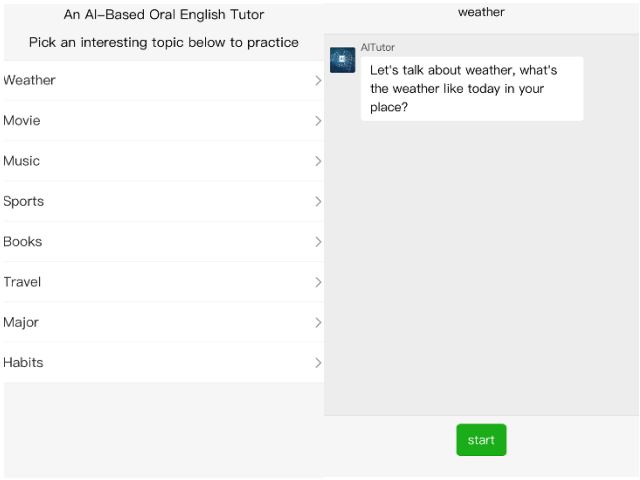

Figure 24: User Interface designed for our Chatbot Level 2 Topic Discussion.

In Figure 25, the level 2 test shows topics without limitation, in general, when the chatbot greets users, the topic are pre-established and dialogue are also limited in this topic. However, since same language model for levels 2 and 3 are used, the discussion after several rounds would exceed the original one and can shows other key-words related to tutor personality in random. At shown in the movie module, the topic would be changed from movie to music at second one.

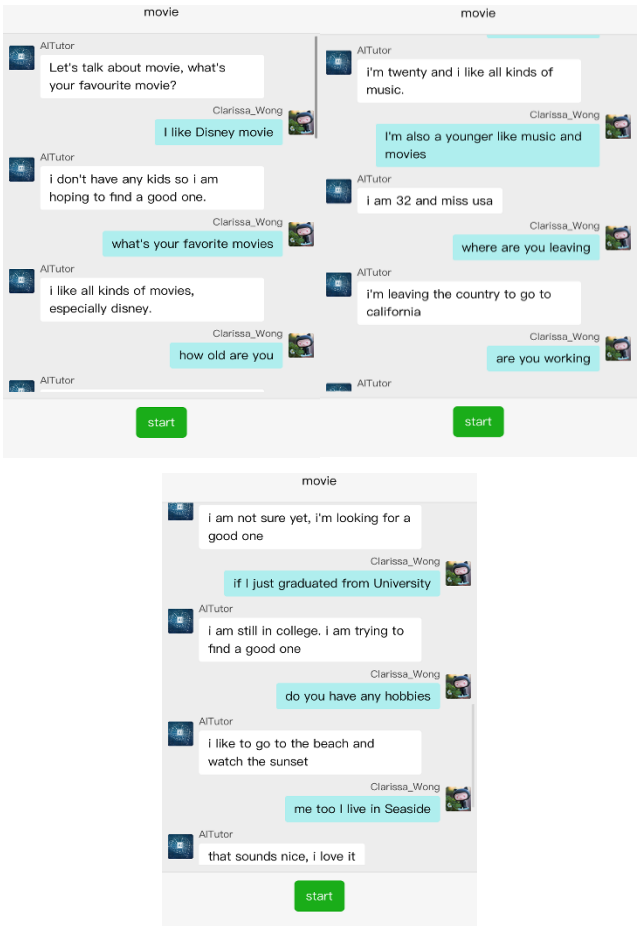

Figure 25: Test of Level 2 Topic Discussion.

At level 3 the same model is used for free chat. To be specific, the system cannot provide sustainable response about the query such as what, why or how. From the visualization of training dataset, the dialogue stayed at information level of the knowledge pyramid. Because the fine-tuned dataset is persona-chat dataset, it cannot provide response like specific *question-answer (QA) system* but prefer to chatbot direction with daily dialogue.

The highest level is free conversation. User and system can have several rounds of dialogue without any restriction . It was found that that the most response is related to fine-tuned dataset contents. Due to the transfer learning model and sequence to sequence model are all data-driven models, with more data used for training, the model has higher language level in syntax in English. Assuming that the text book of an *AI language tutor* is an English teacher, the fine-tuned data will be learned by the machine. After several rounds of test, the tutor uses text book contents to construct sentence content from 200M with high-quality grammar in English by data-driven of 40G text pre-trained.

According to the author of *Transfer Transfo*, the model is equipped with personality of responses generation. During test of Figure 26, every launch of system are different users with different personalities and personal information to chat with users.

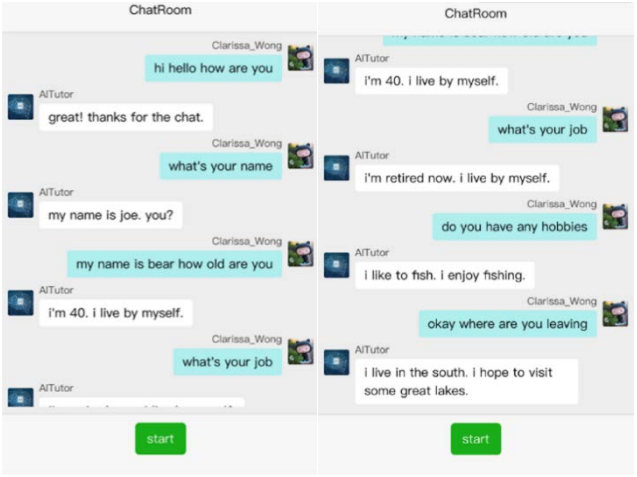

Figure 26: Test of Level 3 Free-Chat with different personality

The daily greetings initially similar to the chat with English teachers in some websites are face to face remote spoken English training. Users cannot guess the next sentence when it is unrelated to any practical topic. Every entry to chatroom has different personality, it is easier to follow the response to continue the subsequent rounds of dialogues.

Different with Level 2 *topic discussion*, Level 3 shows nothing is required to follow in UI. Users could converse anything with chatbot. However, in Figure 27, we can see the response is also limited in the corpus of fine-tune data even though the language model has 40G pre-trained corpus. Level 3 dialogue test results as above showed that English speaking practice with system, the contents cover normally personality with hobbies like music and reading at beginning . When the *AI tutor* ask the key-word 'work', the content will correspond to change and tie to some special words which is related to the key words. Also the contents will be changed from open domain chitchat to another topic proposed by the machine. In this test, "*animal*" and its related words from tutor can also be found at the fine-tune dataset. It was noted that when output consists of semantic of relations amount among different words.

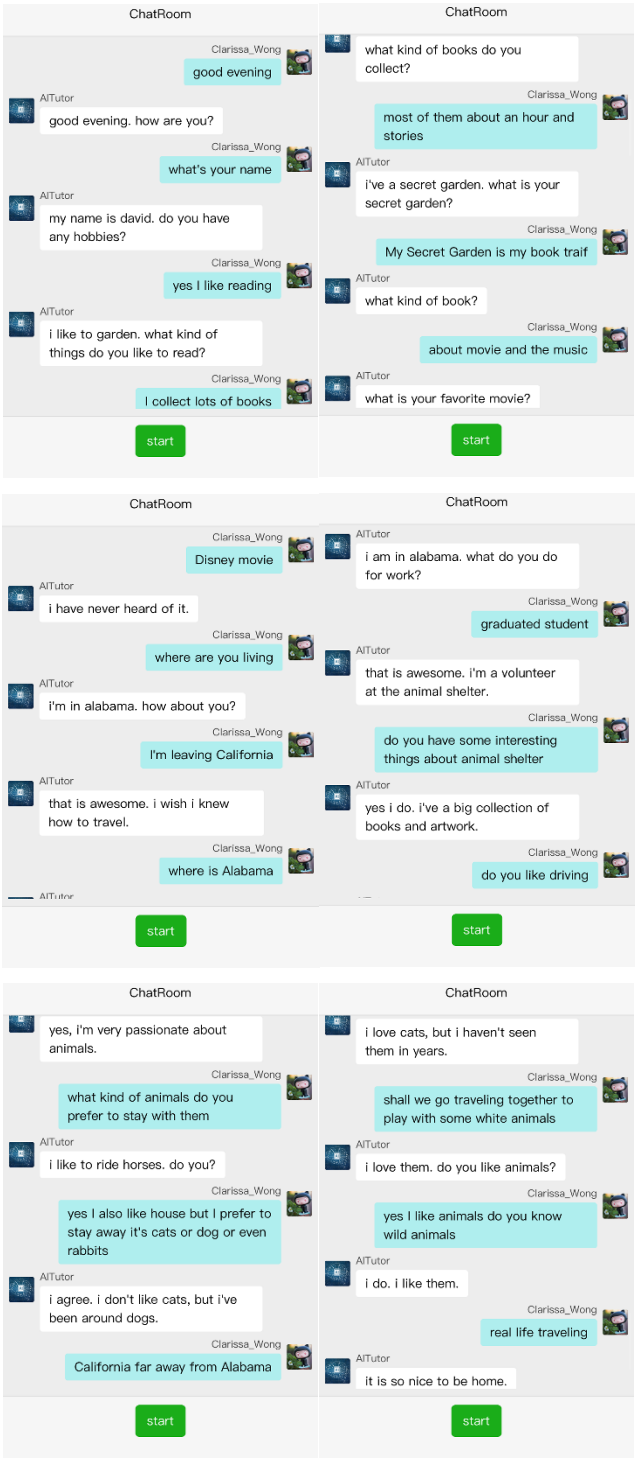

Figure 27: Test of Level 3 Free-Chat

5. Conclusion

In this paper, we used *transfer learning*, an advanced AI technology focusing on technology implementation and system construction to design and implement an *English language chatbot* which is cross-discipline project being proposed the first time and used by real-world university students for learning English round the clock. After around one year improvement, the stable back-end

supports the test with different users 24 hours. Our contribution and the system function shows:

- AI technologies in chatbot field can be well used and combine with the real-world learning needs, not limited in the usage only for personal assistant.
- With magnitude dataset and fine-tune corpus for machine against the AI model, it can generate the meaningful and good syntax response to imitate human response. It overcomes the problem with rule-based or knowledge search engine type chatbot relies on human to prepare the preliminary work, such as label or key-words matching construction. These two methods also range used by ongoing industrial chatbot.

For business aspects, especially in English learning market, there are many applications with splendid *user interface (UI)* and plenty of users declared that their back-end technology uses AI technology. However, during investigation and test commercial apps, the non-mature but with high-expectation neural network model required and relied on dataset quality. For users and system design, the advantages with learning language systematically combines three levels of phonetic, syntactic and semantic which are in both linguistic and the application of *natural language processing* of AI ecosystem.

Based on Turing test theory and *strong AI*, human-machine interaction agent is the main theme of the system. It is proficient in both grammar and knowledge grounds. An Open GPT-2 has already solved the problem to correct error character generated by RNN with pre-trained big data in grammar [3].

6. Future Work

Due to lack of hardware and server, our system performance and capability requires further improvement. At present, 8 GPU with GTX1080Ti on Tencent cloud is used to fine-tune GPT-2 with 345M objective data to make the language model with more meaningful entities to build a QA-based system. However, the computing capacity cannot reach the required magnitude. Also, the official GPT-3 is unavailable for download modification on our GPT-2 language model is continuous until the release of GPT-3. For further research, system optimization for system display is as follows:

- Find a higher capacity server for project implementation
- Optimize UI with better users' experiences and extend the three levels exercises with more options.
- Use 8 V100 GPU to train the largest GPT-2 with existing dataset to compare performance.

Acknowledgment

The authors would like to thank for UIC DST for the provision of computer equipment and facilities. This paper was supported by Research Grant R202008 of Beijing Normal University-Hong Kong Baptist University United International College (UIC) and Key Laboratory for Artificial Intelligence and Multi-Model Data Processing of Department of Education of Guangdong Province.

References

[1] T. Wolf, V. Sanh, J. Chaumond, C. Delangue, "TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents," (ii), 2019.

[2] R.S.T. Lee, Artificial Intelligence in Daily Life, 2020, doi:10.1007/978-981-15-7695-9.

[3] N. Shi, Q. Zeng, R. Lee, "Language Chatbot-The Design and Implementation of English Language Transfer Learning Agent Apps," in 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering, AUTEEE 2020, 403–407, 2020, doi:10.1109/AUTEEE50969.2020.9315567.

[4] Nazar Kvartalnyi, Gamification and Simulation in Education and Corporate Learning, Https://Inoxoft.Com/Gamification-and-Simulation-in-Education-and-Corporate-Learning/, 1, 2020.

[5] M. Qiu, F.L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, W. Chu, "AliMe chat: A sequence to sequence and rerank based chatbot engine," ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), **2**, 498–503, 2017, doi:10.18653/v1/P17-2079.

[6] M. Ghazvininejad, C. Brockett, M.W. Chang, B. Dolan, J. Gao, W.T. Yih, M. Galley, "A knowledge-grounded neural conversation model," 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 5110–5117, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems, **2017-Decem**(Nips), 5999–6009, 2017.

[8] L. Zhou, J. Gao, D. Li, H.Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," Computational Linguistics, **46**(1), 53–93, 2020, doi:10.1162/COLI_a_00368.

[9] C. Agents, "applied sciences Human Annotated Dialogues Dataset for Natural Conversational Agents," 2020.

[10] S. Hoppe, M. Toussaint, "Qgraph-bounded Q-learning: Stabilizing Model-Free Off-Policy Deep Reinforcement Learning," 2020.

[11] D. Ham, J.-G. Lee, Y. Jang, K.-E. Kim, "End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2," **2**(1), 583–592, 2020, doi:10.18653/v1/2020.acl-main.54.

[12] J.B. Tenenbaum, C. Kemp, T.L. Griffiths, N.D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," Science, **331**(6022), 1279–1285, 2011, doi:10.1126/science.1192788.

[13] N. SHI, Q. Zeng, R. Lee, "The Design and Implementation of Language Learning Chatbot with XAI using Ontology and Transfer Learning," (Dm), 305–323, 2020, doi:10.5121/csit.2020.101124.

[14] Jay Alammar, The Illustrated GPT-2 (Visualizing Transformer Langua ge Models), Https://Jalammar.Github.Io/Illustrated-Gpt2/, 1, 2019.