Real-Time Spoken Language Understanding for Orthopedic Clinical Training in Virtual Reality

Han Wei Ng[1], Aiden Koh, Anthea Foong, Jeremy Ong, Jun Hao Tan[2, 3], Eng Tat Khoo[2] and Gabriel Liu[2, 3]

[1] Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore
[2] National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore
[3] National University Health System, 1E Kent Ridge Rd, Singapore 119228, Singapore

Abstract. With the increasing limitation on healthcare resources, lack of real or standardized patients' willingness to participate in medical education and a high medical litigation environment, the traditional learning of "see one" and "do one" is no longer acceptable. Virtual Reality training is becoming commonly used to address this issue. However, some main challenges such as the lack of a dynamic comprehensive conversation and poor language understanding prevents virtual reality from being adopted into mainstream education. In this study, a medical Natural Language Processing (NLP) pipeline is developed. The proposed pipeline displays a total accuracy of 95.6%, with significant improvements compared to the original baseline by an increase of 15% accuracy. The resultant NLP model is subsequently implemented into a medical clinical training virtual simulation.

Keywords: Natural Language Processing, Spoken Language Understanding, Medical Education, Virtual Reality, Virtual Patient.

1 Introduction

Previous research describing the use of technology aims to compensate for students' lack of experience with patients have been reported [1]. The use of chatbots, a text-based intent recognition program, can "impart" knowledge, but does not provide the skills required for a realistic patient-doctor conversation necessary for qualitative medical training. Other researchers have reported preliminary success in non-medical programs using Bidirectional Encoder Representations from Transformers (BERT) [2] mediated natural language processing (NLP) platforms to create conversations. However, this has not been tested in clinical conversations heavily depending on specific medical terminologies. Furthermore, large datasets tend to be biased towards intentions which have more training examples. Augmentation of the dataset can be resource intensive.

Therefore, in this research study, we propose and show that the use of a Dual Intent and Entity Transformer (DIET) NLP model [3] alongside a deep learning sentence augmentation model is able to achieve higher accuracy intention classification in the medical context. The use of synthetic augmentation of sentences serves to increase the amount of training data as well as to ensure a more homogenous dataset, achieving better overall accuracy for a high number of intents.

2 Literature Review

Assistive chatbot technology has seen an increase in uptake over recent years, accelerated by the COVID-19 pandemic. One such use of the assistive chatbot is to deploy chatbots in medical professional training [1]. Medical institutions globally have begun exploring ways to utilize chatbot technology to enhance their current curriculum for training doctors by simulating patients, revising and standardizing examinations. Chatbot technology has also seen much use in training other forms of healthcare professionals such as nurses in terms of patient management [4].

However, chatbot technology remains limited in terms of integration into medical curriculums due to the lack of dynamic conversational capabilities and realism to depict the environment and patient. Since most chatbot systems are founded on rule-based programming, the flow of the conversation and the responses given by the chatbots are linear in nature. This prevents users from being able to feel fully engaged.

Previous work done by Rojowiec et al. showed that the use of Bidirectional Encoder Representations from Transformers (BERT) based models achieved a maximum accuracy of 71.35% on their dataset when tested for intention recognition of the doctor during doctor-patient clinical interviews [5]. It is noted that their collected data shows a high imbalance in the dataset, with a majority of the intent classes having fewer than 20 sample sentences per intent while some intent classes reach above 30.

3 Experimental Methodology

3.1 Dataset Collection

In this study, a medical language corpus is collected from 35 clinical students in their fifth year at the NUS Yong Loo Lin School of Medicine. The medical language corpus consists of questions datasets belonging to questions asked by the doctor and answers given by the patients. In this case, the students are asked to roleplay as doctor and patient pairs. For this study, the disease focused on is an orthopedic disease, spondylolisthesis. To ensure consistency in the medical language corpus, the students are tasked to roleplay as an elderly Chinese female suffering from the disease.

To perform sentence-level intention classification, a total of 212 intention labels were collected and sorted with varying number of sentences per label from the roleplay via voice recording. Transcription was used to convert the audio to text data.

Table 1. Representation of medical auditory data used to train and test the proposed Natural Language Processing model.

Intention Label	Example Sentence 1	Example Sentence 2
Back Pain Location	Could you show me exactly where your back pain is?	Can you point to me where the backache is?
…	…	…
Drug Allergy	Do you have an allergy to any drugs?	Are you allergic to any medication?

4 Sentence-Level Speech Intention Classifier for Orthopedic Clinical Training

4.1 Sentence-to-Sentence Paraphrasing Generator

During the training stage of the NLP model, additional sentences are generated using a pre-trained text-to-text model with Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models (PEGASUS) [6]. A transformer encoder-decoder mode is trained using self-supervised objective Gap Sentences Generation (GSG) for the purpose of generating unique sentences given a single sentence input. The original intention-to-sample ratio spread is shown in Figure 1 (A).

By performing sentence augmentation on the original dataset, we are able to obtain a homogenous spread of sentences across all intention classes as shown in Figure 1 (B). Thus, this reduces the amount of overfitting that the NLP model will have on certain classes. This enables the model to generalize better across all of the identified intent labels, leading to better overall performance of the model.

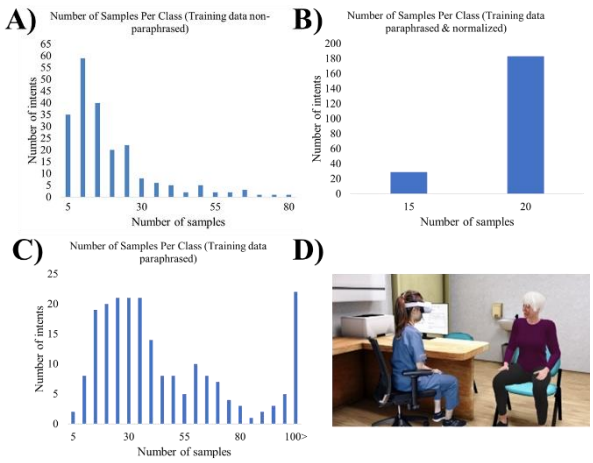

Fig. 1. (A) The original samples per intent distribution. (B) The distribution after performing paraphrasing augmentation. (C) The distribution after normalizing intents to a fixed number of samples. (D) Medical student interaction with virtual patient.

Additionally, it is a well-known issue that audio speech to text services may produce a non-trivial number of incoherent sentences. Thus, the pre-trained paraphraser model can also be used on the input sentences to correct for these mistakes.

The use of the paraphraser model can be seen to be useful in generating new example data using the original sentences as seen in Figure 1 (B) and (C). However, a key limitation behind the paraphraser is its inability to generate a high number of unique examples given a single sentence example. Therefore, the paraphraser has an arbitrary upper limit in terms of the amount of data it can generate depending on the original input sentence. Figure 1 (C) shows the new data spread without homogenizing.

Therefore, the paraphraser is unable to produce a perfectly evenly distributed augmented dataset since this would require producing a high number of augmented sentences to match the intent with the highest number of sample sentences, which is not possible.

4.2 Real-Time Speech Intention Classification

To perform sentence-level intention classification, the open-source DIET NLP model by RASA is implemented. The paraphraser model is used to generate synthetic training data in the medical training context for the DIET NLP model. Subsequently, the DIET NLP is optimized and trained using a combination of permutations involving both the original and generated datasets.

To perform real-time speech intention classification, the audio collected during the testing phase is converted to text using the Azure Speech to Text (Microsoft) service. Transcription error can still occur when converting audio to text. Thus, to reduce the inaccuracy of the trained NLP model, the paraphraser model as described in Section 4.1 is applied to the transcribed text which resolves the grammatical and semantic errors in the sentence while retaining as much of the original meaning as possible.

5 Virtual Reality Clinical Simulation

A virtual reality program for purpose of training a doctor in the clinical setting was created using Unity software as shown in Figure 1 (D). The training simulation aims to mimic a real-life clinical scenario based on a patient who is suffering from an orthopaedic disease, whereby the trainee doctor is tasked to perform differential diagnosis and give a management plan.

The previously trained DIET NLP model gives the conversational capability to the patient avatar. As the DIET NLP model is trained with conversational stories, the appropriate response by the virtual patient is given as output after receiving the speech to text input from the user. If the confidence level of the model given is below that of 50%, the model is tasked to give a null fallback as the output indicating that the model is not certain of the answer. Otherwise, the NLP result which consists of the recognized intent is then collected and checked against a predefined list of actionable intents. The correct patient utterance and animation clip is then selected from a library of recorded audio and animation clips to be played.

6 Results and Discussion

6.1 Overall Accuracy and Ablation Studies

To study the effects of the augmentation as well as the effects of having a more evenly distributed training dataset, an ablation study was carried out.

Including the augmented sentences into the training data yielded a slight increase in the overall F1-score by 0.012. This strongly suggests that the resulting model did not improve significantly due to the inclusion of the paraphrased sentences into the training dataset. The slight improvement can be attributed to the model having a more evenly

distributed dataset and a larger training dataset, encouraging the trained model towards a more generalized optimization.

By applying the paraphraser to only the test sentences, the improvement from the original baseline can be seen to be a significant increase of 0.119 in the F1-score. This is likely due to the paraphraser's model ability to convert the test sentence into a more common form that has a higher probability of having been seen by the model during the training phase, and thus result in higher accuracy.

Table 2. Ablation studies result of the proposed spoken language understanding framework.

Model	Precision	Recall	F1-Score
Rasa Original	0.789	0.844	0.806
+ Paraphrased Training	0.801	0.854	0.818
+ Paraphrased Test	0.917	0.943	0.925
+ Both	**0.950**	**0.967**	**0.956**
+ Both (Normalised)	0.917	0.943	0.925

However, the usefulness of the augmented training data becomes apparent when combining both the training with paraphrased sentences alongside converting the test text input into a paraphrased sentence. This resulted in the highest increase in the model performance with an overall increase in 0.15 in the F1-score to achieve a total score of 0.956. This highly suggests that performing augmentation on both the training and test datasets allow for the mapping of the datasets towards a common domain.

Normalisation of the model did not result in any more increase in performance of the model. However, this is likely to be due to the size of the training dataset having been more than halved due to the forceful removal of sentences to ensure an evenly distributed data. Thus, the difference in the F1-scores suggest that the effect of having a sufficiently large training dataset is more significant as compared to having a more evenly distributed training data.

Therefore, to create the most optimal NLP model for understanding medical-focused intentions we propose the use of a customized paraphraser model that serves to both augment the dataset and to convert the incoming input into a more generalized domain format that is better understanding by the trained model and to reduce sentence errors.

6.2 **Virtual Reality User Study**

To assess the effectiveness of the NLP conversational model for orthopedic clinical training, a total of 23 undergraduate medical students were tasked to interact with the virtual patient. Minimal prompts were given to the students besides an introduction of how to operate the headset and interact with the virtual patient. After going through the entire clinical scenario, the students are then tasked to complete a survey to give their thoughts and opinions on the virtual patient simulator. A copy of the survey and the questions can be found below. The students were asked to rate their experience on a Likert scale from 1 to 5.

19 out of 23 (82.7%) of the students reported a good to excellent overall experience in the clinical training program to fulfil their learning objectives. Sub-group analysis

showed that 69.6% of the students found the clinical interaction with the virtual patient to be similar to their experiences with real-life patient interactions. Furthermore, 86.9% of the student users found the graphics of the program to be realistic. 78% of the student users expressed that the clinical simulation would aid them in the clinical studies.

Overall, the preliminary studies with the undergraduate students strongly indicate a favorable response towards the use of the virtual reality program to enhance their medical clinical training. In addition, the use of the virtual patient simulation can be done at their own leisure, thus aiding in reinforcing the topics learnt and improving memory retention in the students.

7 Conclusion

In conclusion, the study has demonstrated the practical usefulness is the use of a paraphraser sentence augmentation model for the purpose of homogenizing the original training dataset and to convert transcribed audio into a more recognizable form by the model. This offers a significant improvement compared to the baseline, allowing for smoother conversation in a virtual clinical training scenario.

In this study, the trained model captured a total of 212 intents. The importance of training NLP models to recognize a large number of intentions is vital for progress towards building comprehensive interactable virtual agents. Through this study, we have highlighted the feasibility in using the DIET NLP model to perform highly accurate intent recognition on many unique intents. Additionally, the training and testing pipeline proposed by the study has shown that the use of sentence augmentation techniques can greatly improve the model's ability to generalize across testing labels.

References

1. Kaur, A., Singh, S., Chandan, J., Robbins, T., & Patel, V. (2021). Qualitative exploration of digital chatbot use in medical education: A pilot study. DIGITAL HEALTH, 7, 205520762110381.
2. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings Of The 2019 Conference Of The North.
3. Bunk, T., Varshneya, D., Vlasov, V., & Nichol, A. (2020). DIET: Lightweight Language Understanding for Dialogue Systems.
4. Chang, C., Hwang, G., & Gau, M. (2021). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. British Journal Of Educational Technology, 53(1), 171-188.
5. Rojowiec, R., Roth, B., & Fink, M.C. (2020). Intent Recognition in Doctor-Patient Interviews. LREC.
6. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328-11339). PMLR.