IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Towards A Deep Learning Question-Answering Specialized Chatbot for Objective Structured Clinical Examinations

Julia El Zini*, Yara Rizk *, Mariette Awad*, Jumana Antoun†
*Department of Electrical & Computer Engineering
American University of Beirut
{jwe04, yar01, mariette.awad}@aub.edu.lb
†Faculty of Medicine
American University of Beirut
{ja46}@aub.edu.lb

***Abstract*—Medical students undergo exams, called "Objective Structured Clinical Examinations" (OSCEs), to assess their medical competence in clinical tasks. In these OSCEs, a medical student interacts with a standardized patient, asking questions to complete a clinical assessment of the patient's medical case. In real OSCEs, standardized patients or "Actors" are recruited and trained to answer questions about symptoms mentioned in a script designed by the medical examiner. Developing a virtual conversational patient for OSCEs would lead to significant logistical savings. In this work, we develop a deep learning framework to improve the virtual patient's conversational skills. First, deep neural networks learned domain specific word embeddings. Then, long short-term memory networks derived sentence embeddings before a convolutional neural network model selected an answer to a given question from a script. Empirical results on a homegrown corpus showed that this framework outperformed other approaches, and reached an accuracy of 81%.**

***Index Terms*—Natural language processing, convolutional neural networks, LSTM, question answering agent, medical domain semantic understanding, specialized chatbot, OSCE**

I. Introduction

As part of their professional training, medical students must take an exam known as the "Objective Structured Clinical Examination" (OSCE) to test their question asking skills in clinical tasks. The OSCE is a series of stations that simulate various clinical scenarios, one at a time. The medical student interacts with a standardized patient (SP), an actor trained by medical examiners to play the role of a patient, for a very specific medical case. A script instructs the SP on how to answer the medical student's questions, what to say, and more importantly, what not to say. The medical student is allocated a time window to interact with the SP and ask the appropriate questions that are relevant to the diagnosis.

Training actors is an expensive process [1], [2] and a tremendous time investment for medical examiners; one OSCE exam can cost up to $9982 [3]. Actors must be credible and answer questions accurately without revealing too much information. Logistically, multiple stations need to be setup per exam which implies that enough rooms and actors should be reserved. This disrupts normal clinical practice where rooms are used by the examiners to diagnose actual patients. Carrying out OSCEs becomes even more challenging when hundreds of students need to take them.

Beyond facilitating the logistics of administering OSCEs, automating this process also gives medical students access to a wide range of practice tests to improve their skills. Existing virtual SP explored natural language processing (NLP) techniques [4]–[7] to test medical students' social interaction skills [8], medical history, lab test, physical examination and diagnostic skills [9]. Main drawbacks include the narrow range of simulated illnesses or clinical scenarios they simulate and their high error rate.

In this work, we present a VP or specialized chatbot, used interchangeably through the text, for OSCEs capable of interacting with students and responding to their questions based on deep learning. We focus on the NLP engine that allows the avatar to converse with the medical student. Specifically, convolutional neural networks (CNN) and long short-term memory (LSTM) networks learn domain specific word embedding, sentence embedding and answer selection models. This approach does not require explicit linguistic analysis, reducing the burden on system engineers to design suitable features for question-answering (QA) systems. The embeddings model was trained on a corpus of medical documents. An answer selection accuracy of 81% was achieved on a homegrown dataset of QA pairs. Our approach outperformed previous VPs, specific to one pharmacological context, which relied on hand-crafted semantic resources and suffered from high error rates.

The main contributions of this work are: 1) a deep learning framework for answer selection in the medical domain and more specifically OSCEs, 2) domain specific word and sentence embedding models, and 3) a QA corpus for OSCEs. While this work adopts standard deep learning approaches, it applies them to a domain that has received little attention from the machine learning community: OSCEs in medical

This work was funded by the University Research Board (URB) at the American University of Beirut.

978-1-7281-2009-6/$31.00 paper N-20058.pdf

school and more broadly a conversational agent that is required to provide carefully worded answers (not simply correct answers). Also, this domain must also deal with the availability of limited data in the form of few thousand OSCEs as opposed to millions of samples for more general conversational agents.

Next, section II surveys work on VP and deep learning for QA. Section III presents the proposed workflow. Section IV presents the details of the compiled dataset, before section V analyzes the proposed method's empirical performance. Section VI concludes with final remarks.

II. Related work

In this section, we briefly present related work on existing VPs and QA systems, focusing on deep learning approaches.

A. Virtual Patient

Multiple VPs have been proposed in the literature. While some mainly discussed the graphics and animation behind the avatar without detailing the core QA engine [10]–[12], we will focus on those implementations that focused on the QA engine modeling and implementation. Stevens et al. [5] programed a set of answers to specific medical cases into their VP. This restricted the use of the VP to specific predefined medical cases but also resulted in low accuracy (approximately 60%).

Rossen et al. [13] developed a "virtual people factory" where examiners used a graphical user interface to create VPs. Asked a set of questions, the VP answered based on its illness. However, the approach was not very robust: it would easily fail when the questions posed by the medical student slightly differed from those inputted by the medical examiner.

Affective VP [8] tested a medical student's ability to calm agitated patients and their relatives. Although very realistic, the tool had very limited scenarios and required an expert software programmer to configure and modify scenarios. Another VP assessed students' medical history, lab test, physical examinations, and diagnostic skills [5], but received a low student satisfaction rate (6.5/10) due to its high error rate (60%).

Justin, [14] an adolescent male VP suffering from a psychological illness (conduct disorder) frustrated students by its inappropriate answers. Justina, an improvement on Justin, was a female VP suffering from post-traumatic stress disorder [6], [7]. However, it was not very believable, with a high percentage (41%) of "I don't get what you mean" responses due to NLP and speech recognition errors. DIANA [4], [5], a 3D avatar complaining of abdominal pain had a high error rate (40%) in the query handling block.

A recent VP [15] was developed by collecting and analyzing real Chinese medical scenarios. Diagnosis-related concepts were required in this work to construct rules, patterns and features to resolve ellipsis and coreference in the scripts. The VP heavily relied on Chinese semantic featuring which limited the extension to different languages.

In summary, the main weaknesses hindering VPs' effective deployment included their narrow scope (one aspect of diseases), the need for manual configuration or human intervention and the high error rates of the NLP approach. Considering the NLP frameworks of the above-mentioned VPs and others in the literature, they generally relied on language parsers to extract syntactic and semantic information [16]–[18]. Such solutions require human intervention through handcrafted semantic resources.

B. Question-Answering Systems

Two main QA approaches are adopted when the questions are formulated in natural language: open domain QA and restricted domain QA. The former is very narrow and uses specific terminology, whereas the latter is broader in scope.

In both cases, deep learning algorithms have become popular since they shift the burden of syntactic and semantic feature extraction from the system engineer designing sophisticated language parsers [16] to the deep network. For example, CNN, LSTM, and other networks have been adopted in general domain answer selection [19]–[23].

Furthermore, [24] modeled textual compositionality using recurrent neural networks for QA. Yih et al. [25] constructed single-relation QA models using CNN. Yu et al. [19] projected questions and answers into a joint space using deep learning; a vector (hot encoding) reflected a sentence's semantic and syntactic structure. Then, artificial neural networks (ANN), trained on the vector representations, reached a classification accuracy 78% on the TREC dataset [26].

Restricted domain approaches include an insurance-related deep learning QA system [27] where embeddings were obtained using two baseline models: the bag-of-words (BOW) model and an information retrieval model. A multi-layer CNN achieved an accuracy of 65.3%. Yin et al. [28] modeled sentence pairs using an attention-based CNN and achieved 71% mean reciprocal rank. Wang et al. [20] essentially performed keyword matching using bidirectional stacked LSTM for answer selection and achieved a 71% mean average precision and mean reciprocal rank of 79% on the TREC database. To the best of our knowledge, these methodologies have not been explored in the medical domain or for OSCEs, specifically. Unlike other work in the literature, we trained vector representations of words on medical documents instead of relying on pre-trained word embeddings. Thus, performance can be improved by further eliminating ambiguity and leveraging embeddings in a domain specific context. We also incorporated LSTM cells to learn sentence embeddings that capture long-term memory dependencies.

III. Methodology

To converse with a medical student, the VP must understand the script provided by the medical examiner, parse the medical student's question and select the correct answer(s) from the script. Next, we present the workflow that allows the VP to perform these tasks.

A. Virtual Patient Workflow

The chatbot's overall design, shown in Fig. 1, consists of the following modules. Speech recognition transforms the student's speech into text, whereas speech synthesis transforms

the NLP engine's text output into speech. The graphics module includes a renderer and an animator to display and animate the 3D avatar (Fig. 2) and lip sync during audio playback. The NLP engine, word embeddings and scripts database - enclosed in red rectangles and the focus of this work - allow the VP to converse with the medical student. An answer selection approach is adopted over an answer generation system. Medical examiners want the VP to answer with carefully worded responses to faithfully model OSCEs: VP should not bias the assessment and diagnosis of the medical student by leading them to believe they have a different medical condition. Next, we describe our NLP engine by formulating the answer selection problem and solving it in a deep learning framework.

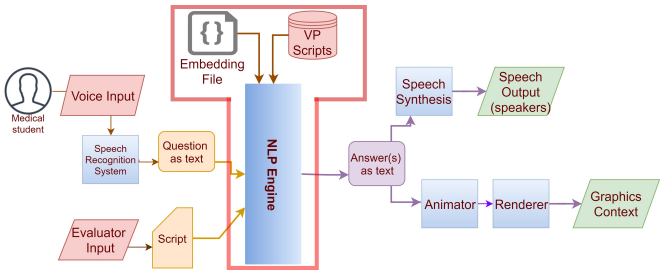

Fig. 1. VP Workflow

Fig. 2. A Screenshot of our VP Avatar

B. NLP Engine

The goal of the NLP engine is to correctly answer questions based on a script. We chose to adopt an answer selection model because, in the context of OSCEs, it is not only important to answer questions correctly but to carefully word the answer as to avoid misleading the student and hence unfairly influencing the outcome of the exam. Furthermore, we adopt a supervised learning scheme to learn a mapping between QA pairs and a judgment of whether the pair is a correct match. Unsupervised learning is a challenge in the OSCE domain because data is not abundantly available to learn correct patterns and cluster the data.

The script, provided by the medical examiner at the start of each OSCE session, is viewed by the VP as a set of m candidate answers. Asked a question q, the VP compares it against each sentence $s_i, 1 \leq i \leq m$ and outputs a judgment y_i for each sentence s_i, where $y_i = 1$ if s_i is a correct answer for question q and 0 otherwise. In the case where $y_i = 0$ for all $1 \leq i \leq m$, i.e. the VP could not find an answer to the question in the given script, it returns a special "no answer" output. Thus, our task is to learn a classifier over triples of the form (q, s_i, y_i) to allow the VP to predict the correct judgment over QA pairs. The classifier works as follows: upon receiving a question, the question and each candidate answer are fed to two networks that compute their word embeddings, as shown in Fig. 3. The outputs of each networks are then fed to a new network (LSTM) that computes the sentence embedding. A deep learner (CNN in our case) operates on the QA sentence embeddings and outputs a judgment for each pair (q, s_i).

As shown in Fig. 3, this is achieved by: 1) training a word embedding model on medical documents to obtain domain specific word embeddings, 2) learning sentence embedding of the QA pairs using LSTM networks, and 3) training a classifiers over the QA embeddings using CNN networks.

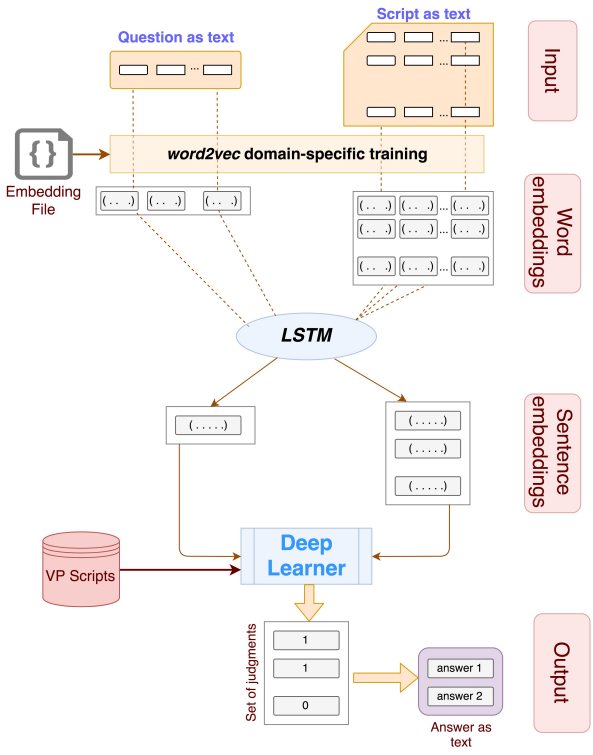

Fig. 3. Answer Selection Workflow

1) Domain Specific Word Embedding: Training deep learners on NLP tasks requires the conversion of words to numerical vectors. Vector space models embed words in a continuous vector space, while accounting for their semantic and syntactic information by mapping similar words to nearby points. *Word2vec* is a widely used predictive model for learning word embeddings. Existing approaches in QA use the *word2vec* word vector model pre-trained on Google News corpus. Pre-trained models have the advantage of being trained on a huge dataset, but they may not perform well in domain specific settings. Training *word2vec* on medical documents may result in a more accurate disambiguation of words since embeddings would reflect the meaning in a medical context. This domain-

specific pretraining falls under the transfer learning research area. For instance, the knowledge about word syntax and semantics within the medical domain is transferred from an unsupervised learning model (domain-specific word2vec) to the supervised QA model. In this work, a word embedding model is trained on a dataset of full-text biomedical journal archives.

2) Sentence Embedding Model: Having transformed words into vectors, these vectors can be used to train deep learners. Since our application handles sentences, we can either represent these sentences by their constituting words' vectors or combine them to form a vector representing the sentence and capturing complex structure of sentences. Previous approaches [19] relied on two famous models to generate sentence embeddings from word embeddings: the BOW model (shown in Fig. 4) and the Bi-grams model (shown in Fig. 5). While the former represents a sentence by the average of the vectors of its constituting words, the latter encodes more semantic features of a sentence using a CNN where one convolutional layer builds bi-grams (pairs of consecutive words) and the pooling layer captures the semantic information of the full sentence. Although this model accounts for word ordering, it fails to capture long-term memory dependencies in a sentence.

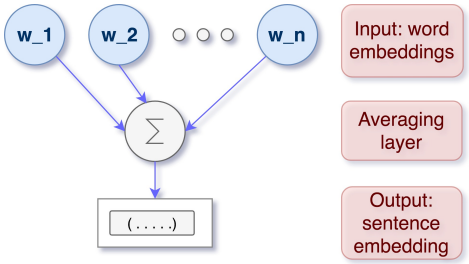

Fig. 4. BOW model architecture

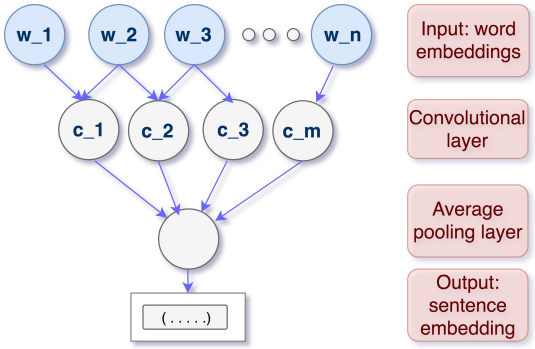

Fig. 5. Bi-gram-based ANN architecture

Both models fail to capture long dependencies in a sentence. BOW cannot to account for word ordering and does not capture more complex semantic structures of sentences. Bi-grams fail to capture long-term memory dependencies in a sentence. To address these difficulties, LSTM cells [29] were incorporated into our model. LSTM sequentially takes each word in a sentence and embeds its information into a semantic vector, as shown in Fig. 6. As it processes the sentence, LSTM cells accumulate richer information. Reaching the last word, the hidden layer of the network will hold the semantic representation of the whole sentence [30].

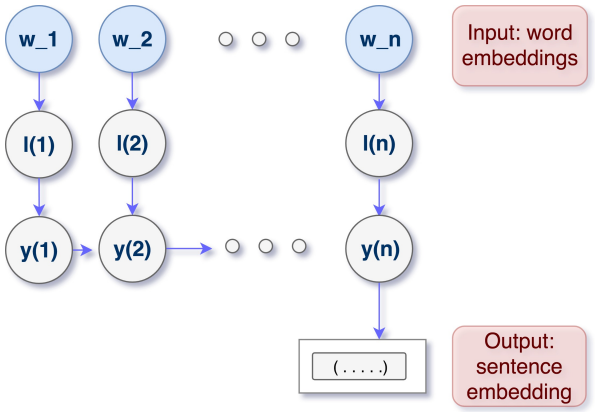

Fig. 6. LSTM-based sentence embeddings

3) QA Classification: Once a question is received and its embedding vector is computed, a CNN operates on each pair (q, s_i) and outputs the corresponding judgment. The CNN is trained on a dataset manually derived from scripts previously used in OSCEs and described next.

IV. Dataset Details

The NLP engine relies on two different datasets: the first is to build the domain specific word embeddings and the other is to train the QA classifier.

A. Word Embeddings Dataset

The word embedding model [1] was trained on a subset of full-text biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine [31]. The dataset consisted of 28,000 documents with 8.8 million distinct sentences and 1.5 billion words. Out-of-vocabulary words were mostly medication names that did not occur while training. They were assigned an UNKNOWN token that mapped to a specific word embedding vector.

B. QA Dataset

1) Dataset Generation: A labeled dataset[2] was created to train our answer selection models in a supervised framework. Given a script, we manually generated corresponding questions for every sentence, treating the QA pair as having a positive judgment $y = 1$. Then, pairs with judgment $y = 0$ were generated by randomly assigning a sentence from the script to a question that did not correspond to it. The ratio $(y = 1) : (y = 0)$ was $1 : 3$; this choice was based on faithfully representing real world scenarios while avoiding training complications from severely unbalanced data. Different balance ratios were investigated and the corresponding

[1] Available on GitLab: gitlab.com/awadailab/medical-specific-embeddings
[2] Available on Kaggle: kaggle.com/awadailab/vp-for-qa

performances are reported in Section V. Moreover, for a given sentence (candidate answer) in the script, multiple questions were generated in-house by medical examiners and an NLP researcher to account for the different ways a question can be asked. For example, regarding dizziness, the script provides the following information: "I don't really feel dizzy". The medical student, however, may not ask the direct question "Do you feel dizzy?" but "Do objects usually spin around you?" or "Have you recently felt lightheaded?". Thus, multiple variants of the same question were included in the training data to train a more robust model with a richer dataset. Moreover, some questions include ellipses or casual wording to mimic actual conversations during the medical assessment process.

2) Dataset Description: The QA dataset consisted of 4,332 triplets of the form (q, s_i, y_i), derived from more than 50 scripts provided by the Faculty of Medicine at the American University of Beirut[3] and previously used in OSCEs. The scripts (Fig. 7) correspond to a wide variety of medicine branches: cardiology, neurology, endocrinology, orthology, nephrology, and others. On average, we had 14 sentences per script (with 5 as minimum and 43 as maximum), 178 words per script (with 37 as minimum and 690 as maximum) and 13 words per sentence (with 3 as minimum as 48 as maximum). The preprocessing phase consisted of converting scripts to QA pairs (Table I), as described in the previous section IV-B1.

Table II reports the complexity and readability of the generated questions according to metrics that are widely used in the medical context [32]. The Flesch reading ease formula [33] outputs a score ranging from 0 to 100 with 100 being the score of a "Very easy to read" text. The score is computed as in (1) and the reported score, 72, falls in the category of a "Fairly-easy-to-read" text.

$$\text{Flesch reading ease score} = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right) \quad (1)$$

The other metrics [33]–[36] estimate the years of education needed for a person to understand the text. The flesch Kincaid grade level [33] computes the score as:

$$\text{Flesch grade level} = 0.39 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \quad (2)$$

Whereas, the SMOG index [34], which is non-linear in the number of sentences, is computed as:

$$\text{grade} = 1.0430\sqrt{\text{nbr of polysyllables} \times \frac{30}{\text{nbr of sentences}}} + 31291 \quad (3)$$

The DaleChall readability score [36] and the Gunning FOG index [35] are computed as in (4) and (5) respectively.

[3]The OSCEs can be downloaded from https://mededportal.com/

$$\text{Dale-Chall score} = 0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right) \quad (4)$$

$$\text{SMOG} = 0.4\left(\frac{\text{words}}{\text{sentences}} + 100\left[\frac{\text{complex words}}{\text{words}}\right]\right) \quad (5)$$

The reported scores show that elementary/ intermediate education is needed to understand the scripts on the first reading. The FOG index significantly differs from other indeces and shows that a higher education level is needed to understand the text. The same tests were applied on the generated questions as shown in Table III. The results show that the standard deviation of the questions readability is consistently higher than that of the scripts reflecting thus the diversity of the question types used in the generation process. Since the SMOG index needs a text of at least thirty sentences, it was not applicable in the questions readability tests and only scripts with more than thirty sentences were considered in the scripts readability tests.

I am a previously healthy 27 year old woman. I have been having chest discomfort about twice a week for the last 2 weeks. It is sharp, associated with difficulty getting a deep breath. It seems to come on mostly at work or when I am driving. It lasts about a half hour at a time. I have tried Tylenol, Advil, drinking cold water, and antacids without much benefit. It does not radiate. It is severe enough to interrupt my work but not excruciating. I have not had any heartburn or stomach symptoms. I am concerned that it could be a heart problem.
I smoke 15 cigarettes a day. I am trying to quit; I had cut down from 1 pack/day to ½ a pack but recently went back up to ¾ of a pack, probably from stress. No drug use. I have never been in the hospital or been told I had any chronic illnesses, never had anything like this before, never had a cholesterol test. I am separated from my husband of two years, which is very stressful. We had argued a lot and just grown apart, with no history of domestic violence. I work as a bank supervisor. I do not have children. I am not currently sexually active or using birth control. I do aerobics 3 to 4 times a week. I have not had any problems with chest pain or breathing while exercising; actually that is when I feel best. I take a multivitamin daily with no other meds. I have no allergies.
My father had a heart attack last year when he was 64. This is one of the reasons why I am worried about these pains. He also smokes and has high blood pressure. My mother and older brother are healthy.

Fig. 7. Sample OSCE script

TABLE I
QA SAMPLES

A:	I can get dressed and go to the bathroom myself
Q:	Are you able to take care of yourself?
A:	I try not to eat sweets.
Q:	Do you watch your diet?
A:	My father has high blood pressure.
Q:	Any health problems in your family?
A:	The chest pain came on suddenly whilst I was in bed
Q:	When did you first feel the pain?

V. Experimental Results

A. Experimental setup

The experiments were run on an Intel Xeon 64-bit 12-core processor machine with a Quadro K2000 NVIDIA GPU.

TABLE II
SCRIPTS READABILITY

Metric	Average ± stdev	Min	Max
Flesch Reading Ease [33]	72 ± 11	31	89
Flesch Kincaid Grade [33]	6 ± 2	12	3
SMOG Index [34]	9.3 ± 1.4	6	12
Dale-Chall readability [36]	6.4 ± 3	0.15	16
Gunning FOG Index [35]	14 ± 2.3	7	20

TABLE III
QUESTIONS READABILITY

Metric	Average ± stdev	Min	Max
Flesch Reading Ease [33]	82 ± 29	120	7
Flesch Kincaid Grade [33]	3 ± 4	18	0.2
SMOG Index [34]	NA	NA	NA
Dale-Chall readability [36]	6 ± 4	0.15	14
Gunning FOG Index [35]	12 ± 6	29.8	3.2

The algorithms were written in Python 2.7 using the Keras and Tensorflow packages. A 5-fold cross fold validation was adopted in all experiments and the reported accuracies were averaged over all folds. The best results of a grid search to find the optimal hyper-parameters of support vector machines (SVM) [37], extreme learning machines (ELM) [38], [39], ANN and CNN are reported. A CNN with 2 convolutional layers (1st layer: 300 64x64 filters, 2nd layer: 64 10x10 filters) and one maxpooling layer was trained using 30% dropout. In this work, we assume the input is already in text form, i.e. we do not utilize the end to end avatar described in section III-A. This means that the reported accuracies are not affected by the error rate of the speech-to-text module; they are solely from the NLP model.

B. Answer Selection Performance

A direct comparison with previous VP approaches was not possible because their corpus is not available publicly, and their approaches relied on engineered semantic resources, unlike our approach. Instead, we experimented with different embeddings and classifiers. To highlight the task's difficulty level, we compare the best model against two baselines: a weak baseline that randomly selects an answer from the script and a strong baseline that retrieves from the script the closest sentence as an answer to the question. In the latter, sentences that have the smallest Euclidean distance between their embedding vectors are considered close. Since the number of answers to each question is not predefined in advance, we assume that all the questions have only one answer in the script and we calculate the accuracy as: $\frac{\text{Questions correctly answered}}{\text{Total questions}}$. The results in Table IV show that both baselines have a poor performance compared to our model. The weak baseline performance is inversely proportional to the number of sentences in the script. Whereas, the strong baseline excessively simplifies the QA task and models it as a Euclidean distance between the embeddings reaching thus an underfitting representation of the training data.

TABLE IV
BEST MODEL VERSUS BASELINES ACCURACY

Best Model	Weak Baseline	Strong Baseline
83%	7%	52%

TABLE V
DIFFERENT BALANCE RATIOS FOR THE BEST ARCHITECTURE

Balance ratio	Accuracy (%)	F-measure (%)
1:1	76	68
1:2	78	71
1:3	81	76
1:4	83	75

Table VI reports the performance of various answer selection models for a balance ratio of 1:3. "Medical40" indicates that the word embedding model was trained on a subset (40% or approximately 2.7GB) of the medical document dataset, whereas "Medical100" refers to the model trained on the entire dataset (approximately 5.6GB). A "None" for sentence embedding means that the word embeddings were directly fed to CNN without further computing sentence embeddings. BOW and bi-grams (BiGr) sentence embeddings models were also compared to our LSTM sentence embeddings. The best model, with an accuracy of 81%, was achieved by training the word embeddings on the full medical dataset, deriving sentence embeddings using LSTM then using a CNN to select answers. On average, LSTM outperformed other approaches by 8%. Furthermore, CNN, which had three layers of 64 sigmoid units each and used a dropout of 30%, outperformed other classifiers when fixing word and sentence embedding models by an average of 4%. The best architecture was tested on different balance ratios, and the results are reported in Table V. The table shows that the unbalanced dataset (1:4 ratio) resulted in a better accuracy than the perfectly balanced one (1:1 ratio). The best accuracy (83%) was achieved with a 1:4 ratio with a slightly worse f-measure, 75% compared to 74% achieved with a 1:3 balance ratio. This improved performance of unbalanced datasets can be explained by the fact that they contain more samples than the balanced ones as a result of our dataset generation technique, explained in Section IV-B1. In fact, with more negative samples generated for each positive QA pair, the size of the dataset increases which negatively impacts the balance ratio of the data but provides the models with more training samples.

1) Word Embedding Model: The domain specific model's answer selection accuracy was consistently better than the pretrained model (almost 18% on average based on Table VI). Table VII provides an example of a QA pair that was incorrectly associated when using the pretrained word embedding model but correctly associated by our domain specific model; the latter led to the correct association of "paracetamol" and "ibuprofen" with "pain killers" and the recognition of "basal cell carcinoma" as a medical illness. Fig. 8 clearly illustrates this fact by plotting the principle component analysis (PCA) reduced word vectors of "paracetamol" and "ibuprofen". The

TABLE VI
ANSWER SELECTION ACCURACY (%)

Embedding		Classifier			
Word	Sentence	CNN	ANN	SVM	ELM
Pretrained	None	53	51	51	52
	BOW	56	53	54	53
	BiGr	56	54	53	54
	LSTM	64	62	53	58
Medical40	None	64	59	52	58
	BOW	68	62	53	61
	BiGr	70	69	54	62
	LSTM	76	69	54	59
Medical100	None	76	70	69	65
	BOW	75	73	74	72
	BiGr	75	74	74	73
	LSTM	**81**	73	75	73

distance between both words when represent by vectors from the pretrained model is over 3 times greater than that of model trained on medical data (3.42 vs. 0.98). Therefore, domain specific embedding models are essential to the success of VP's answer selection.

TABLE VII
CORRECTLY ASSOCIATED QA SAMPLES, WHEN WORD EMBEDDINGS ARE TRAINED ON MEDICAL DATA

Q:	Do you have any medical history?
A:	I had a basal cell carcinoma removed from my face six months ago.
Q:	Do you take any pain killers?
A:	It is a severe ache, but paracetamol, ibuprofen and a heat pack have helped.

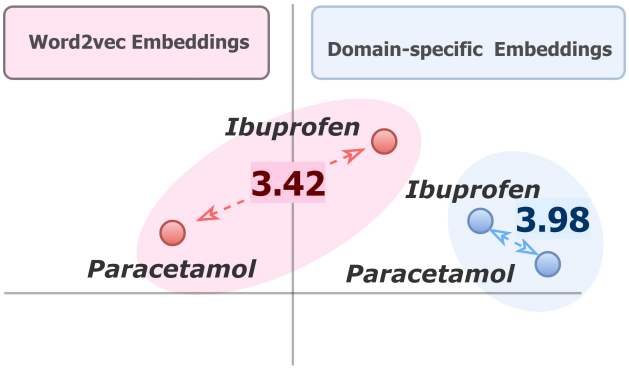

Fig. 8. PCA-reduced word embeddings

Moreover, we compared the accuracy for various word embedding dimensions, as shown in Table VIII. A dimensionality of 300 achieved the best accuracy while requiring less training time but requiring more memory to store the model.

2) Word Embeddings Model Size: Increasing the training set size improved the performance of the domain specific word embedding model, as shown in Table VI, but it affected the size of the model. Training on the entire medical document corpus resulted in a model that required 40GB of space, whereas training on only 40% required 18GB. Also, training time increased exponentially as the number of processed

TABLE VIII
ANSWER SELECTION ACCURACY (%) WHEN VARYING THE WORD EMBEDDING DIMENSION

Embedding		Dimension		
Word	Sentence	64	100	300
Medical40	None	60	60	64
Medical40	LSTM	71	72	76

documents increased, since updating the embedding model required more operations as words were passed through the network. The network processed almost 5 million words per second on the first few documents but only about 500 words per second on the last few, requiring a total of over four weeks to train. Since our model is trained offline and accuracy is crucial to our application, we opted to incur the additional computational costs.

3) Sentence Embeddings: Table VI clearly shows that deriving sentence embeddings by training an LSTM model outperforms BOW and bi-grams models. Table IX shows a sample QA pair that was incorrectly associated when sentence embeddings were not used but correctly associated when they were. Training on medical documents allowed the network to associate "vision" with "visual acuity" and "bumping into things", leading to a correct classification.

TABLE IX
INCORRECTLY ASSOCIATED QA SAMPLE WHEN SENTENCE EMBEDDINGS WERE NOT USED

Q:	Any vision problems?
A:	Recently I seem to be bumping into things, such as door handles at home or walls at work, but I feel that my visual acuity is okay and I do not feel my balance is reduced.

TABLE X
CONFUSION MATRIX REPORTS THE PER CLASS ACCURACIES (%) OF THE BEST MODEL

	Actual: Yes	Actual: No
Predicted: Yes	30	14
Predicted: No	5	51

4) Statistical Significance Tests: We performed the Friedman and ANOVA statistical tests on the various models we presented. A p-value of 7.5e-10 and 0.0295 for each test respectively which are less than 0.05 and imply that there is a statistical difference between the models. We also ran the test for pairs of models. For example, for Medical100-LSTM-CNN and Medical100-None-CNN both Friedman and ANOVA tests showed no statistical difference between both. However, this does not necessarily imply that on a larger test set, both models will still be equivalent.

5) Classification Error Analysis: Table X presents the confusion matrix of the best model, which consists of a word embedding model with dimension 300 trained on the full medical dataset, LSTM for sentence embeddings and CNN for answer selection. The false positive rate is 14% and the false

negative rate is 5%. Since our system could return multiple sentences as answers to a question, false positives imply that the VP is providing additional information that the medical student did not ask about. On the other hand, false negatives imply that the VP is omitting information. We consider false positives more harmful since in that case, the medical student would not ask relevant questions that the VP already gave answers to. Thus, the student would receive a lower grade during the evaluation phase. Therefore, it is important that our system achieves a low false positive rate.

Focusing on some of the VP's mistakes, the first two samples shown in Table XI are false positives and the fourth is a false negative. In the first error, the VP model correctly associated "pain" with "headaches" but was misled by the association between "when" in the question and "few hours" in the answer. However, this answer should be selected when a question like "How long does the pain last?" is asked. In the second error, "previous" and "last few months" caused a high correlation between the question and answer, in addition to "trauma" and "headaches". According to the script, this answer conveys information about the current issue and "previous trauma" refers to other issues. The third error was due to the high semantic association between smoking and drinking; this is a clear example of providing additional information to the medical student. If the medical student didn't ask about drinking habits, s/he would be penalized in the evaluation. Finally, the VP agent could not classify the fourth QA sample as a match since the model did not associate "suffering from" with the inability to bend.

TABLE XI
INCORRECTLY ASSOCIATED QA SAMPLES

Q:	When did you first feel the pain?
A:	The headaches tend to last a few hours at a time.
Q:	Any previous trauma?
A:	I am presenting to the GP with the complaint of worsening headaches over the last few months.
Q:	How often do you smoke?
A:	I drink alcohol on the weekend, totaling 16 units per week.
Q:	What are you suffering from?
A:	I find it difficult to bend.

6) Script-level Case Study: Next, we analyze the VP's performance on the overall script rather than the QA pairs. The scripts used for testing have different difficulty levels according to the Flesch Kincaid ease formula. Easy, Plain and Hard have score ranges of 70-100, 50-69 and 30-49, respectively. We rank the answers selected by the VP according to the probability of their retrieval (as outputted by the classifier): if the probability of retrieval is above a certain threshold the answer is selected. We report the results for different thresholds: 0.3, 0.5 and 0.7.

Table XII reports three rates for each readability level. CorrAnsOut is the average percentage of the questions (across scripts), whose correct answers are selected (possibly with incorrect answers too). CorrAnsOutFirst is the average percentage of the questions whose correct answers were selected with higher ranks. CorrAnsOutOnly is the average percentage of the questions whose correct answers **only** were selected (i.e. without false positive).

The difficulty level of the script affected VP's performance. For example, "I always had tension type headaches like most people now and again, but this feels different: it has followed a rather busy period with work, due to recent school examinations, with late nights drinking coffee whilst marking assessments.", is a sentence in a "Hard" script that the VP failed to associate with the question: "When do you mostly feel your headache?" Since the CorrAnsOut and CorrAnsOutFirst rates are very close then when the VP outputs true negatives, it outputs them with low ranks. When the threshold is high (0.7 i.e. only answers with 70% confidence are selected), the corrAnsOutOnly rate becomes close to the CorrAnsOut which means that the false negative rate has decreased.

TABLE XII
DIFFERENT METRICS FOR THE ACCURACY ACROSS SCRIPTS

Threshold	Metric	Easy	Plain	Hard
NA	Number of Scripts	8	5	2
0.3	CorrAnsOut	92	89	81
	CorrAnsOutFirst	89	87	79
	CorrAnsOutOnly	61	62	51
0.5	CorrAnsOut	89	87	79
	CorrAnsOutFirst	86	85	74
	CorrAnsOutOnly	75	72	64
0.7	CorrAnsOut	86	84	70
	CorrAnsOutFirst	85	82	67
	CorrAnsOutOnly	85	81	65

VI. CONCLUSION

OSCEs, an integral but expensive part of medical school, test students' question asking skills in clinical scenarios. To reduce their cost and simplify their logistics, we proposed a conversational avatar to simulate a medical patient actor. The best model achieved an 81% accuracy by obtaining word embeddings from a model trained on medical documents, sentence embeddings from an LSTM trained on the word embeddings and a CNN trained on a homegrown corpus of QA pairs. Future work will expand the corpus to explore different answer generation and sentence embedding methods.

ACKNOWLEDGMENT

We would like to thank the Maroun Semaan Faculty of Engineering and Architecture and University Research Board at the American University of Beirut for the funding of this project.

REFERENCES

[1] R. M. Harden, P. Lilley, and P. Madalena, *The definitive guide to the OSCE: the objective structured clinical examination as a performance assessment*. Elsevier, 2016.

[2] C. B. Lee, L. Madrazo, U. Khan, T. Thangarasa, M. McConnell, and K. Khamisa, "A student-initiated objective structured clinical examination as a sustainable cost-effective learning experience," *Medical edu. online*, vol. 23, no. 1, p. 1440111, 2018.

[3] T. Rau, J. Fegert, and H. Liebhardt, "How high are the personnel costs for osce? a financial report on management aspects," *GMS Zeitschrift für medizinische Ausbildung*, vol. 28, no. 1, 2011.

[4] R. Dickerson *et al.*, "Evaluating a script-based approach for simulating patient-doctor interaction," in *Int. Conf. Human-Computer Interface Advances for Modeling & Simulation*, vol. 1, 2005, pp. 79–84.

[5] A. Stevens *et al.*, "The use of virtual patients to teach medical students history taking and communication skills," *The American J. Surgery*, vol. 191, no. 6, pp. 806–811, 2006.

[6] P. Kenny, T. D. Parsons, J. Gratch, and A. A. Rizzo, "Evaluation of justina: a virtual patient with ptsd," in *Int. Workshop on Intelligent Virtual Agents*. Springer, 2008, pp. 394–408.

[7] T. Parsons, P. Kenny, and A. Rizzo, "Virtual human patients for training of clinical interview and communication skills," in *Virtual Reality & Associated Technology Conf.*, 2008.

[8] B. Jung, A. Ahad, and M. Weber, "The affective virtual patient: an e-learning tool for social interaction training within the medical field," in *Proc. Int. Conf. TESI-Training Education & Education*, 2005, pp. 1–9.

[9] O. Courteille *et al.*, "The use of a virtual patient case in an osce-based exam–a pilot study," *Medical Teacher*, vol. 30, no. 3, pp. e66–e76, 2008.

[10] N. Zary, G. Johnson, J. Boberg, and U. Fors, "Development, implementation and pilot evaluation of a web-based virtual patient case simulation environment–web-sp," *BMC medical edu.*, vol. 6, no. 1, p. 10, 2006.

[11] K. Maicher *et al.*, "Developing a conversational virtual standardized patient to enable students to practice history-taking skills," *Simulation in Healthcare*, vol. 12, no. 2, pp. 124–131, 2017.

[12] N. Rombauts, "Patients virtuels: pédagogie, état de lart et développement du simulateur alphadiag," Ph.D. dissertation, Thèse de médecine, Faculté de Médecine, Univ. Claude Bernard, Lyon, France, 2014.

[13] B. Rossen and B. Lok, "A crowdsourcing method to develop virtual human conversational agents," *Int. J. Human-Computer Studies*, vol. 70, no. 4, pp. 301–319, 2012.

[14] P. Kenny, T. D. Parsons, J. Gratch, A. Leuski, and A. A. Rizzo, "Virtual patients for clinical therapist skills training," in *Int. Workshop on Intelligent Virtual Agents*. Springer, 2007, pp. 197–210.

[15] C.-J. Lin, C.-W. Pao, Y.-H. Chen, C.-T. Liu, and H.-H. Hsu, "Ellipsis and coreference resolution in a computerized virtual patient dialogue system," *J. Medical Systems*, vol. 40, no. 9, p. 206, 2016.

[16] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *EMNLP*, vol. 2, no. 5, 2013, p. 6.

[17] Q. Cai and A. Yates, "Large-scale semantic parsing via schema matching and lexicon extension," in *ACL*, 2013, pp. 423–433.

[18] T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer, "Scaling semantic parsers with on-the-fly ontology matching," in *EMNLP*. Association for Computational Linguistics (ACL), 2013.

[19] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, "Deep learning for answer sentence selection," *arXiv preprint arXiv:1412.1632*, 2014.

[20] D. Wang and E. Nyberg, "A long short-term memory model for answer sentence selection in question answering," in *Proc. 53rd Annu. Meeting of ACL & 7th IJCNLP*, vol. 2, 2015, pp. 707–712.

[21] Y. Yang, W.-t. Yih, and C. Meek, "Wikiqa: a challenge dataset for open-domain question answering," in *EMNLP*, 2015, pp. 2013–2018.

[22] X. Zhou *et al.*, "Icrc-hit: a deep learning based comment sequence labeling system for answer selection challenge," in *Proc. 9th Int. Workshop on Semantic Evaluation*, 2015, pp. 210–214.

[23] M. Tan, C. dos Santos, B. Xiang, and B. Zhou, "Improved representation learning for question answer matching," in *Proc. 54th Annu. Meeting of ACL*, vol. 1, 2016, pp. 464–473.

[24] M. Iyyer, J. L. Boyd-Graber, L. M. B. Claudino, R. Socher, and H. Daumé III, "A neural network for factoid question answering over paragraphs," in *EMNLP*, 2014, pp. 633–644.

[25] W.-t. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proc. 52nd Annu. Meeting of ACL (Volume 2: Short Papers)*, vol. 2, 2014, pp. 643–648.

[26] M. Wang, N. A. Smith, and T. Mitamura, "What is the jeopardy model? a quasi-synchronous grammar for qa," in *EMNLP-CoNLL*, vol. 7, 2007, pp. 22–32.

[27] M. Feng, B. Xiang, M. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: A study and an open task," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2015, pp. 813–820.

[28] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: attention-based convolutional neural network for modeling sentence pairs," *Trans. of ACL*, vol. 4, no. 1, pp. 259–272, 2016.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] H. Palangi *et al.*, "Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 4, pp. 694–707, 2016.

[31] C. Ré and C. Zhang, "DeepDive open datasets," http://deepdive.stanford.edu/opendata, 2015.

[32] D. Wu *et al.*, "Applying multiple methods to assess the readability of a large corpus of medical documents," *Studies in health technology & informatics*, vol. 192, p. 647, 2013.

[33] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, Univ. of Central Florida, 1975.

[34] G. H. Mc Laughlin, "Smog grading-a new readability formula," *J. Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[35] R. Gunning, *The technique of clear writing*. McGraw-Hill, NY, 1952.

[36] J. S. Chall and E. Dale, *Readability revisited: the new Dale-Chall readability formula*. Brookline Books, 1995.

[37] M. Awad and R. Khanna, *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress, 2015.

[38] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *IJCNN*, vol. 2. IEEE, 2004, pp. 985–990.

[39] Y. Rizk and M. Awad, "On extreme learning machines in sequential and time series prediction: A non-iterative and approximate training algorithm for recurrent neural networks," *Neurocomputing*, vol. 325, pp. 1–19, 2019.