Investigate Predictive AAC with Sequence to Sequence Network on Small Conversational Datasets

Yuehan Pei

May 5, 2022

Abstract

An Augmentative and Alternative Communication (AAC) device is a tablet or laptop that helps someone with a speech or language impairment to communicate. Making good question-answer predictions can help accelerate the communication. In this project, I investigate RNN, GRU, LSTM under Encoder Decoder Network, with both regular decoder and attention decoder. The Word Error Rate shows that Sequence to Sequence Network, especially Attention RNN, which is trained on small conversational data sets, can provide promising results on AAC device's question-answer pair predictions.

1 Introduction

AAC stands for Alternative and Augmentative Communication, people with speaking disorders often use AAC devices (Figure 1) to communicate with doctors and caregivers. Common forms of speaking disorders include stuttering, cluttering, Down Syndrome, apraxia, dysarthria, aphasia, Parkinson's disease, amyotrophic lateral sclerosis (ALS), or cerebral palsy. Currently there are three different approaches being applied on AAC devices. One of them is to let users select words or images from its interface and form a sentence, the speaking partner can read it or hear it via text-to-speech. Since many patients with speaking disorder also have injuries, surgeries, cognitive or physical development problems, their typing speed is very slow, often less than 10 words per minute. [Higginbotham et al., 2007] [Simpson et al., 2006] [Trnka et al., 2009] So predictive AAC devices normally use a language model to calculate and make suggestions of likely upcoming text. The next-word predictions are based on the content that the user has just typed in. The model does not consider the other side of the conversation, it can't be personalized by users' own dialogues, either.

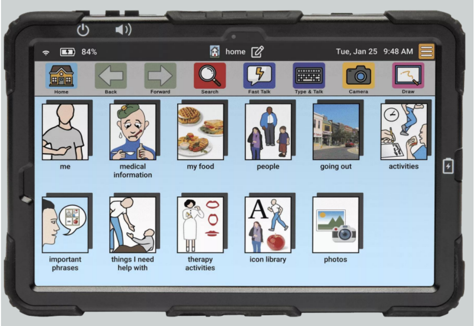


Figure 1: A typical AAC device

To better serve the users, in this paper, instead of making suggestions for just one next word, I make predictive AAC which generates a whole sentence as an answer based on the speaking side of the conversation, and this model can be personalized by the users with a very small dialogue data set of their own.

2 Related Work

Currently there are three different approaches being applied on AAC devices.

First one is to transform the original disordered speech signal in the spectral domain by using a cycle-consistent Generative Adversarial Network (CycleGAN), and to synthesize a new speech signal from the trained model [Yang and Chung, 2020]. They give reading prompts to the patients to get parallel audio data and use GAN to train the samples. After they get the synthesized speech signal, it is evaluated via speech-to-text to calculate word error rate (WER). The evaluation shows that cycle-consistent adversarial training is a promising approach for dysarthric speech conversion tasks. Similar works include using voice conversion (VC) to improve the speech intelligibility of surgical patients who have had parts of their articulators removed [Chen et al., 2018]. This

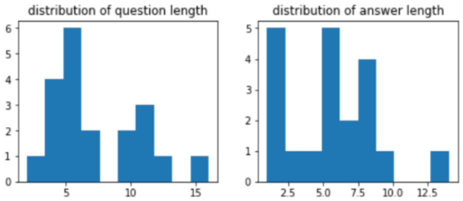


Figure 2: The distribution of sentence lengths.

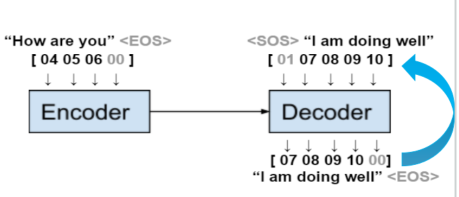


Figure 3: The Encoder Decoder Architecture

is the first end-to-end GAN-based unsupervised VC model applied to impaired speech.

The second approach is to apply automatic speech recognition (ASR) on the impaired speech. IBM investigate a variety of distributed deep learning strategies for ASR and evaluate them with a state-of-the-art Long short-term memory (LSTM) acoustic model. Their system trains a model to WER 7.6% on Switchboard and WER 13.1% on CallHome in less than 12 hours. This is the fastest system that trains these tasks to this level of accuracy [Zhang et al., 2019]. Google also investigates the performance of personalized ASR for recognizing disordered speech using small amounts of per-speaker adaptation data. They trained personalized models for 195 individuals with different types and severities of speech impairment with training sets ranging in size from ¡1 minute to 18-20 minutes of speech data. WER thresholds were selected to determine Success Percentage (the percentage of personalized models reaching the target WER) in different application scenarios. For the home automation scenario, 79% of speakers reached the target WER with 18-20 minutes of speech; but even with only 3-4 minutes of speech, 63% of speakers reached the target WER. Their results demonstrate that with only a few minutes of recordings, individuals with disordered speech could benefit from personalized ASR. [Tobin and Tomanek, 2022]

3 Data Analysis

The first step is to obtain question-answer-pair text data reasonably representative of everyday person-to-person conversations. The dataset is found in this chatterbot corpus. Many of the files are not daily language, so I only pick the file named ”conversations”. The total number of question-answer pairs is 109. The distribution of sentence length is in Figure 2. Most of them are very short.

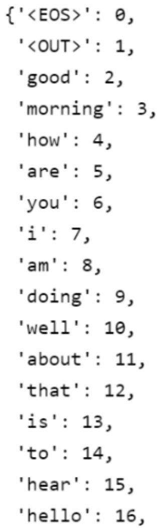
I split the data into train set and test set. Train Dataset is used to fit the machine learning model. Test Dataset is used to evaluate the fit machine learning model. I put 89 question-answer pairs into Train Dataset and 20 into Test Dataset.

4 Language Modeling Experiments

Predictive AAC devices typically use an N-gram language model (LM). An N-gram LM calculates the probability of a token given the previous N-1 tokens. The performance of this model depends on the training data being closely matched to a user's text. [Adhikary et al., 2019]

Recently, recurrent neural network language models (RNNLMs) have achieved state-of-the-art performance over traditional N-gram language models. RNNLMs have been shown to better model long range dependencies when combined with techniques such as long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] or gated recurrent units (GRU) [Chung et al., 2014].

A Recurrent Neural Network, or RNN, is a network that operates on a sequence and uses its own output as input for subsequent steps. A Sequence to Sequence network, or seq2seq network, or Encoder Decoder network, is a model consisting of two RNNs called the encoder and decoder. The encoder reads an input sequence and outputs a single vector, for every input word the encoder outputs a vector and a hidden state, and uses the hidden state for the next input word. The decoder reads that vector to produce an output sequence. (Figure 3) The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Unlike sequence prediction with a single RNN, where every input corresponds to an output, the seq2seq model frees us from sequence length and order, which makes it suitable for generating answers based on questions. After 89 question-answer pairs are read into the model, the encoder vocabulary size is 210, the decoder vocabulary size is 243. Figure 4 shows a sample of the encoded vocabularies.



```
{'<EOS>': 0,
 '<OUT>': 1,
 'good': 2,
 'morning': 3,
 'how': 4,
 'are': 5,
 'you': 6,
 'i': 7,
 'am': 8,
 'doing': 9,
 'well': 10,
 'about': 11,
 'that': 12,
 'is': 13,
 'to': 14,
 'hear': 15,
 'hello': 16,
```

Figure 4: Every word is encoded

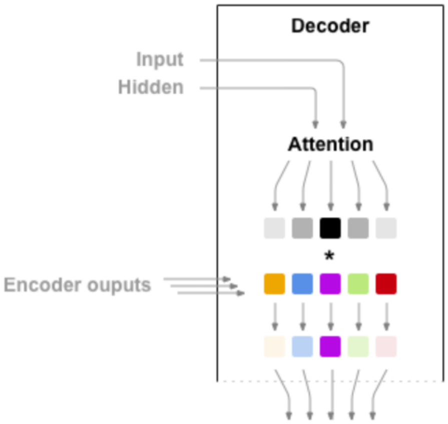


Figure 5: The Attention Decoder

If only the context vector is passed between the encoder and decoder, that single vector carries the burden of encoding the entire sentence. Attention allows the decoder network to "focus" on a different part of the encoder's outputs for every step of the decoder's own outputs. (Figure 5) First we calculate a set of attention weights. These will be multiplied by the encoder output vectors to create a weighted combination. The result (called attn_applied in the code) should contain information about that specific part of the input sequence, and thus help the decoder choose the right output words. [Zeineldeen et al., 2021]

In this paper, I investigate six different models under Encoder Decoder network: Recurrent Neural Network, Gated Recurrent Unit, Long Short-Term Memory, Attention RNN, Attention GRU, and Attention LSTM.

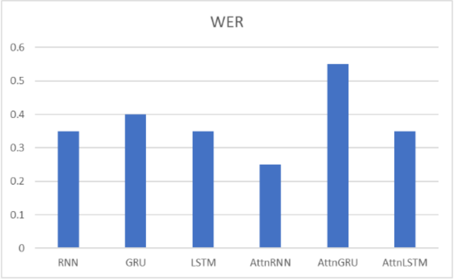


Figure 6: Results of the six models

5 Results

After the experiments of the six models, I got their word error rates in Figure 6. Attention-based RNN achieves the lowest Word Error Rate as 25%.

Since we're using a very small dataset, the WER might change dramatically if we use another new dataset. It is uncertain to determine which model will work the best on each user's personalized dataset. So the best way is to get all the WERs and compare them, then apply the model which has lowest WER to that specific customer.

6 Conclusion

AAC users often face challenges in taking part in everyday conversations due to their underdevelopment. Predictions can provide an opportunity to accelerate their communication rate, but it is crucial these predictions be as accurate as possible. Leveraging real-world contextual clues offers one route to improving these predictions. This paper provides results showing Sequence to Sequence Network, especially Attention RNN, which is trained on small conversational Datasets, can provide promising results on AAC device's question-answer pair predictions.

7 Future Work

Future work will be focused on the following directions:

- Make the conversational dataset more strongly related to the user's life, for example, add dialogues about their health conditions.

- Try different models to continue to improve prediction accuracy.

- Perform speech recognition using speech recognizers, then collect our conversation data via speech-to-text.

- Redefine WER to handle special cases when the prediction has no same word as the correct answer but has the same meaning. For example, the question is ”how are you”, the correct answer is ”I am doing well”, but if the prediction was ”fine, thank you, and you?”, there are no same words with the correct answer, the WER will be 100%. But it is a good answer. We need to find ways to deal with these situations.

References

[Adhikary et al., 2019] Adhikary, J., Watling, R., Fletcher, C., Stanage, A., and Vertanen, K. (2019). Investigating speech recognition for improving predictive aac. In *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies.*

[Chen et al., 2018] Chen, L.-W., Lee, H.-Y., and Tsao, Y. (2018). Generative adversarial networks for unpaired voice transformation on impaired speech. *arXiv preprint arXiv:1810.12656.*

[Chung et al., 2014] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555.*

[Higginbotham et al., 2007] Higginbotham, D. J., Shane, H., Russell, S., and Caves, K. (2007). Access to aac: Present, past, and future. *Augmentative and alternative communication*, 23(3):243–257.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Simpson et al., 2006] Simpson, R., Koester, H., and LoPresti, E. (2006). Evaluation of an adaptive row/column scanning system. *Technology and disability*, 18(3):127–138.

[Tobin and Tomanek, 2022] Tobin, J. and Tomanek, K. (2022). Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6637–6641. IEEE.

[Trnka et al., 2009] Trnka, K., McCaw, J., Yarrington, D., McCoy, K. F., and Pennington, C. (2009). User interaction with word prediction: The effects of prediction quality. *ACM Transactions on Accessible Computing (TACCESS)*, 1(3):1–34.

[Yang and Chung, 2020] Yang, S. H. and Chung, M. (2020). Improving dysarthric speech intelligibility using cycle-consistent adversarial training. *arXiv preprint arXiv:2001.04260.*

[Zeineldeen et al., 2021] Zeineldeen, M., Glushko, A., Michel, W., Zeyer, A., Schlüter, R., and Ney, H. (2021). Investigating methods to improve language model integration for attention-based encoder-decoder asr models. *arXiv preprint arXiv:2104.05544.*

[Zhang et al., 2019] Zhang, W., Cui, X., Finkler, U., Kingsbury, B., Saon, G., Kung, D., and Picheny, M. (2019). Distributed deep learning strategies for automatic speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5706–5710. IEEE.