Guyriano Charles

DREU 2022

Introduction

The purpose of this paper is to describe the research undertaken during my DREU experience from May 2022 to August 2022 at the University of Alabama. This research attempts to solve two problems. It attempts to ameliorate the lack of African American dialectal speech in the training and testing of automatic speech recognition systems (ASRs) while also providing an easier way for African Americans in rural communities to find medical or otherwise related services. The research impacts ASR research as well as social justice for African Americans who use African American Vernacular English (AAVE) and those with Alzheimer's and related dementias (A&RDs). African Americans and those with A&RDs are most interested in this research. The current solution is to wait until the major ASR developers such as Google and Amazon find some reason to target AAVE communities and there does not appear to be any incentive to do so currently given the success of ASR systems used in products such as Echo and the Google Home aid. We propose a diverse group of engineers and ASR developers as well as the help of corpuses such as the CORAAL corpus to find well formatted data on AAVE. This research is different by taking a hard emphasis on AAVE and the broader connections between medical infrastructure and speech.

Related Work

Approach 1:

The first approach comes from Markl and Lai, 2021. The approach to the issue in this paper is to document predictive bias in ASRs while considering user experience, socio-historical and sociolinguistic context, harms (re)produced by the system, and technical aspects of ASR. Methodologies and knowledge drawn from include HCI, sociolinguistics, research on fairness in AI, and SLT. Intersectional benchmarks also needed to be established to connect multiple demographic axes linked to interlocking structures of oppression (e.g. race and gender) cannot be considered separately.

This paper is mainly a case study of The Lothian Diary project which is an ongoing interdisciplinary research project inviting residents of the Lothians region of Scotland to contribute self-recorded audio and video diaries about their experiences of the COVID-19 pandemic. More than 120 diaries have been collected as of the publication of the paper. They are highly variable in recording quality, number of speakers and topics discussed, and participants' age, gender, linguistic background, ethnicity, socio-economic class, and level of education. There is also a wide range of other first and second language varieties of English, as well as other languages. The Lothian Diary project also includes many of these other varieties of English, rather than focusing on speakers with long residential histories in a particular area or first language speakers.

The following are assumptions from this study: It is not clear that the intended or current user base is reflective of all use cases or potential users, it is possible that significant variation in performance between user groups is hidden by reporting an average across all tested recordings. Multiple demographic axes linked to interlocking structures of oppression cannot be considered separately. It is thus important to create intersectional benchmarks. It is also assumed that WER does not account for the context or effect of an error. Understanding the context of errors is useful since errors are both more likely to occur and to be severe in particular phonetic, prosodic and lexical contexts. It is assumed that evaluations of SLT systems rarely reflect explicitly on how users interact with them. Finally, the societal context that an ASR system is developed in and implemented allows us to identify the specific harms it could inflict on users see the underlying societal structures giving rise to predictive bias. Identifying risk and causes in turn allows us to mitigate harms and bias in the future.

The results were that word error rate (WER) for individual speakers varies dramatically Some of these errors seemed to be related to accent differences. For example, Scottish speakers' pronunciations of I or I've are frequently mis-transcribed as ah or of among other accent-based errors. There was also significant variation within each accent group. GC STT failed to transcribe filled pauses (uh, um) and word fragments and occasionally deleted false starts and repetitions. Errors also appeared to be more prevalent in the vicinity of hesitations and repetitions. Therefore, more hesitant, and repetitive speakers tend to have higher error rates. The highest WERs in this sample also tended to be present in speakers who used a lot of Scots words.

Approach 2:

The second approach to be discussed comes Trinh and Droppo et. al, 2022. The researchers utilize the elastic weight consolidation (EWC) regularization loss to identify directions in parameters space along which the ASR weights can vary to improve for high-error regions, while still maintaining performance on the speaker population overall. The results demonstrate that EWC can reduce the WER in the region with highest WER by 3.2% relative while reducing the overall WER by 1.3% relative. The researchers also evaluate language and acoustic models' role in ASR fairness. They also propose a clustering algorithm to identify WER disparities based on geographic region. The main assumption here is that initial parameters of the model learned in initial training tend to be lost as new data is acquired. Their approach is to use an ASR system with an end-to-end model utilizing the recurrent neural network transducer (RNN-T) architecture which calculates acoustic and semantic features of input speech simultaneously. The system includes a long short-term memory (LSTM) encoder, an LSTM prediction network, and a joint network. EWC can be applied to make the model acquire new knowledge while retaining information already learned. A continual learning setup for ASR is used where the researchers adapt a pretrained model (trained on data from all regions) on new data that contains only speech from regions having high WER, without access to previous data. A clustering tree algorithm is used to split the dataset into different subsets, where each subset corresponds to a specific geographic region by approximate longitude and latitude. The algorithm clusters similar data into their own subsets.

The training dataset is comprised of 47,000 hours of audio split into several subsets. The ASR pretraining set is created by drawing randomly 10,000 hours of speech from the corpus. The proposed method is compared against other transfer learning techniques in eight experiments. In Experiment 1, a baseline ASR model is trained. In Experiment 2, the researchers adapt the pretrained model on 10,000 hours of utterances with decreasing WER from all regions without freezing any parameters. The encoder parameters are frozen in Experiment 3 while the predictor parameters are frozen in Experiment 4. Next, the parameters of the first three layers of the encoders and the first layer of the predictor are frozen. EWC regularization starts in Experiment 6. In Experiment 7, an ASR is trained from scratch, rather than adapting a pretrained model. In Experiment 8, the ASR is trained from scratch with a dataset that is a combination of pre-training and training data.

The results of this paper focus on reducing geographic differences in ASR performance, but their method is equally applicable to other scenarios with a need to adapt a model to a specific dataset without degrading overall performance.
The results show that the best result overall is obtained when the ASR parameters are not frozen (Exp. 2). In that case, the region with the highest WER is
reduced by 2.9% compared to the baseline, while freezing the encoder, or only its first three layers and the first predictor layer, reduces the region WER max by 1.4% and 2.5%, respectively. Freezing the encoder and predictor keeps new information from being taken in by the model so it seems to be a general poor way to reduce WER.

Approach 3:

The third approach comes from Nam and Lake, 2022. This approach is based on two recently collected corpora of conversational speech. The first is the Corpus of Regional African American Language (CORAAL), a collection of sociolinguistic interviews with dozens of black individuals who speak African American Vernacular English (AAVE) to varying degrees. Language models underlying commercial ASR systems are not readily available. It is likely, however, that these systems use language models that have similar statistical properties to models that are publicly available, like Transformer-XL , GPT, and GPT-2. Disparities between these three models can be compared using the publicly available versions that have been pretrained on large corpora of text data. The standard performance metric for language models is perplexity, which roughly can be viewed as the number of reasonable continuations of a phrase under the model. Perplexity is computed under the GPT-2 language model, as well as the GPT-1 and Transformer-XL models and disparities are compared between results from the CORAAL data and data of Standard English. Under all three language models, we find the average perplexity of snippets by black speakers is lower. This means better performance than the average perplexity of snippets by white speakers using the CORAAL sample data. Transformer-XL has perplexity of 115 for black speakers compared with 153 for white speakers. GPT has perplexity of 52 and 68 for black and white speakers, respectively; and GPT-2 has perplexity of 45 and 55, respectively. This is interesting since when examining the lack of the word “be” in the African Americans’ speech where it would normally be in Standard American English, perplexity increased for the CORAAL data.

The researchers believe this difference is at least partially due to the relative number of unique words spoken by black and white sample members. Although the total duration and number of words spoken by black and white speakers in our sample were similar, black speakers uttered fewer unique words (5,651) than white speakers (6,280).

Guyriano Charles

DREU 2022

All else being equal, a smaller vocabulary generally yields lower model perplexity, as it is easier to predict the next word in a sequence. This means that the lexical and grammatical properties of ASR systems do not account for the large overall racial disparities in WERs. Since these snippets from black speakers have fewer unique words and lower perplexity, they should be easier for the ASRs to transcribe. These results suggest that the problem may instead lie with the acoustic models underlying ASRs. The same is true for short phrases in terms of perplexity. These results suggest that racial disparities in ASR performance are related to differences in pronunciation and prosody including rhythm, pitch, syllable accenting, vowel duration, and lenition between white and black speakers.

One limitation of the study is that the audio samples of white and black speakers came from different geographical areas of the country. It is possible that some of the differences are a product of regional linguistic variation. It is noted by the researchers that there are two reasons to believe that AAVE speech itself is driving the results. First, word error rate is strongly associated with AAVE dialect density. Second, the two California sites of white speakers that were included, Sacramento and Humboldt, exhibited similar error rates despite diversity in regional speech patterns across the state and differences in the socio-geographical contexts of these two locations. Humboldt is a rural community, whereas Sacramento is the state capitol.

REU Approach

Some unaddressed issues with the above papers are that they do not all attack systemic and intersectional nature of racial bias in ASRs. They also do not all broadly apply Elastic weight consolidation. The new approach devised in my REU research is to deploy speech models trained on CORAAL data in a practical setting. Ours is in helping Alzheimer's and Related Dementia (A/RDs) afflicted people. This manner of deployment helps to normalize the use of AAVE speech by providing a specific context for using such data. Hopefully, this would provide incentive as well as encourage competition among commercial ASR developers to develop similar systems. This new approach was inspired by Dr. Monica Anderson and graduate student Adria Mason going out into rural Alabama communities (e.g., Pikkins County) and realizing typical ASR models do not work well with AAVE speech. We foresee that the new approach will work because it extensively maps the needs of those who use it via a community asset mapping backend. Such a backend is designed to hold onto and provide invaluable data into the needs of people using the system. This data would be coming in very frequently if deployment goes well so it would also be less difficult and more practical to retrain the model in the future or solely use new data. These are already two standard ways to retrain ASR models. We are therefore highly invested in being ethical with this data, for example, following one major tenet of ethical computing where data such as location is not held past its window of use.

We are using Rasa and Kaldi. The new software consists of Kaldi scripts designed to work with data from any source by automating downloading the data, parsing it and also accounting for pauses and non-speech vocalizations such as steupses in the data which typical ASRs tend to not to. Following all this, the model is built, trained and tested automatically as well. Kaldi is an ASR development toolkit. Rasa is a learning framework for building AI assistants and chatbots. I worked on the Kaldi side of the project.

The following code is from the getdata.sh Bash script in the CORAAL directory. This script is designed to download and parse audio data which is what the following code sample does. It also separates this data into training and testing data.

```
    for text in *.txt; do
        $(sed -i "1d" $text) #remove the column names from the text and only leave the actual data
        #Extract all lines from the text files that are not the interviewer and removes brackets and braces or parentheses and slashes from around action descriptors.
        $(awk -F"[*][(*)<*>/*/]" '$0 !~ /'int'/ { print $3, $NF }' $text > awk_tout.txt)
       $(awk '$0 ~ /'VLD_se0'/ && $0 !~ "[(]" && $0 !~ "[[]" && $0 !~ "[<]" && $0 !~ "[/]" { print $1, $3, $NF }' awk_tout.txt > awk_out.txt) || exit
        for i in *.wav; do
            if [[ ${i##*.}==${text##*.} ]]; then
                python3 parse_wav.py "$i" "awk_out.txt" #segment the lines from above out of the interview wav file
                mv ./*sub*.wav ./audio #the segments have a descriptor "sub" in their filename, moves these to the audio folder.
            fi
        done
        cp $text data/train #move the text file to both the train and test folders
        mv $text data/test
    done
```

Guyriano Charles

DREU 2022

The following snippet is from the run.sh script. This snippet is meant to create multiple metadata files (spk2gender, utt2spk, wav.scp, text) for the data and move it into the training and testing directories.

```
while read -r line; do
    ran=$(tr -dc A-Za-z0-9 </dev/urandom | head -c 6)
    ranstr="-"$ran
    spkID=${line%.*}
    ind_pre=${spkID##*b}
    ind=$((ind_pre-1))
    ext=".""${line##*.}"
    uttID=$spkID$ranstr
    echo $spkID ${line:12:1} >> spk2gender
    echo $uttID $spkID >> utt2spk
    meta=$(find . -type f -name ${spkID%_sub*}".txt") #remove the sub index from the file name first to match it to a meta text file
    python3 loc/get_trans.py $meta $ind $uttID 0
    echo $uttID $(pwd)'/'$uttID$ext >> wav.scp
    cd ..
    cd ..
    mv audio/$line audio/$uttID$ext || exit 1
    mv audio/$uttID$ext data/test
    cd data/test
done < $loctmp/speakers_test.txt
```

These new algorithms are compared to algorithms in the Voxforge directory in Kaldi. Dr. Monica Anderson, my REU mentor, ran the Voxforge scripts as a test of what my code's functionality may look like. We hypothesize that word error rate when using the CORAAL speech corpus will be lower when using speech data from rural Alabama communities. We will test it with CORAAL data but also using a chatbot for users to interact with. Initially, we are testing the chatbot and ASR model amongst ourselves as we use AAVE to varying degrees and this project is targeted at such people. Over the course of the next year, data from the chatbot will be compiled and used to retrain the model using Raza, Kaldi, as well as the Alabama state supercomputer. Phone (sound) data as well as WER data. Spoken data from the chatbot will be saved in conversation documents that Rasa compiles as the chatbot is used. Audio data can also be saved on any recording devices as .wav files with a 44100 kHz sample rate. The experiments outlined above would add more data as well as these referenced in those papers. We predict that among the A/RDs population, one phenomenon that may not be apparent by looking at the data is perplexity between technical speech describing medical services, tools, etc. and normal casual speech. Casual speech is statistically more likely to include words that are much more popular in the English language than technical speech so more technical speech may have a higher WER.

Conclusion

In all, the lack of AAVE data and speech corpuses in training ASRs is the focus of this paper. We hypothesize that an ASR model trained on AAVE data in a practical setting will lead to lower WER and lower perplexity than demonstrated by commercial ASR systems. Initial results aligned with our hypothesis. Implications are that users of AAVE in rural southern communities as well as their caretakers will be more independent with the community asset mapping utility developed from this research. The results are generalizable to rural Alabama and perhaps the south more in general. AAVE, however, is not the same in all regions of the United States so that must be taken into account. In the future, we may test out our asset mapping utility in different regions of the USA e.g. northeast, southwest and west coast.

Guyriano Charles

DREU 2022

Works Cited:

Koenecke, Allison, et al. "Racial disparities in automated speech recognition." *Proceedings of the National Academy of Sciences* 117.14 (2020): 7684-7689.

Markl, Nina, and Catherine Lai. "Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation." *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. 2021.

Trinh, Viet Anh, et al. "Reducing Geographic Disparities in Automatic Speech Recognition via Elastic Weight Consolidation." *arXiv preprint arXiv:2207.07850* (2022).