Research Seminar, Extensive review, 16/07/2021
Learning to Understand Child-directed and Adult-directed Speech
Ernst Hamer, s1047547
Josip Koprcina, s1062758

Summary

In the paper of Lieke Gelderloos et al (2020) the differences between learning from child directed speech (CDS) and adult directed speech (ADS) are explored. CDS differs from CDS mainly in acoustic and linguistic aspects. The acoustic aspects include slow speech rate, exaggerated intonation, and higher pitch. The linguistic aspects include short sentences, repetition of words and phrases, word choice, and that words appear in isolation more often. In the paper it was found that in the initial stages of learning with the SBERT algorithm child directed speech is more helpful but in the end they have comparable results to ADS. On the other hand, the model trained on ADS generalizes better. In addition they found that the differences in learning are mostly due to linguistic properties of the speech data and not the acoustic. The authors found this by also training on synthetic speech.

1. Objective of the paper

The objective of the paper is to improve extracting meaning from speech by automatic speech recognition systems. In the paper, the authors match speech to a semantic representation. The authors explore the differences between learning from CDS and ADS and model the process from speech to meaning. In this process, acoustic and higher level factors are separated by the use of a synthetic speech dataset for experiments. The paper claims that, Although with CDS learning is initially faster, in the end ADS generalized better. The authors claim this to be due to linguistic rather than acoustic properties, because of the experiments on synthetic data. The overall research question of the paper is as follows; is child directed speech easier to learn to understand than adult directed speech? When learning to map speech to semantic information with machine learning models.

2. Evidence Given

a. Data

The NewmanRatner corpus (Newman et al., 2016) was used as a dataset. This is a collection of speech from free play sessions between caregivers and children aged 7 to 24 months and caregivers interviewed by researchers. The reason why this dataset is used in particular is because it features the same people first talking towards children, of course using CDS, then speaking with interviewers using ADS. This means they have both types of speech from the same origin. In the paper, recall was used as an evaluation metric. The authors did 3 runs of the algorithm per speech set (CDS, ADS, synthetic CDS, synthetic ADS). Speech and

Sentence-BERT embeddings were mapped based on the model used in Merkx et al., 2019. The characteristic statistics of the CDS and ADS datasets are displayed in Table 1. In Table 1 it can be seen that the vocabulary size and the total number of words is significantly larger in ADS than in CDS. In addition the words per utterance is more than doubled in ADS compared to CDS. This of course makes sense, when we talk to a child we elongate our words thus using less per sentence but because of the difference in pronunciation length both types take about the same amount of time. The CDS dataset ended up having more instances and because of that, the 21,465 instances were taken out randomly out of the ADS dataset to keep a class balance. 1000 segments were taken out of both ADS and CDS for both validation and test sets.

Dataset	**CDS**	**ADS**
Vocabulary size	3,170	5,665
Total nr. of words	97,118	203,084
Type/token ratio	.033	.028
Words per utterance	4.52	9.46
Utterance length in seconds	3.37	3.46
Words per second	1.34	2.74

Table 1: Descriptive statistics of the data

As for the synthetic dataset, it was created by giving the above mentioned instances of CDS and ADS to the Google text2speech API which pronounced all of them the same way. This way any differences in speech were removed. That also meant the CDS segments now lasted about half as long as ADS ones because of the lower number of words per utterance. The synthetic data is a lot cleaner simply because it’s automatically created and the natural speech data was collected using a microphone attached to the speaker's clothes which resulted in worse quality of audio. This quality was the same for CDS and ADC so it shouldn’t have influenced the results significantly.

b. Method

In the paper, Speech and Sentence-BERT embeddings are mapped based on Merkx et al., 2019. The speech is downsampled by a convolutional layer and is fed into 4 bidirectional GRU layers, followed by an attention operator. This method for semantic embeddings using linear mapping to joined space (single linear transformation) is decoding. The novelty in their approach was the use of SBERT sentence embeddings instead of visual vectors. The model works by making the cosine distance between similar words smaller compared to words that are not as similar.

c. Results

The models are evaluated on recalls of 1, 5 and 10, meaning whether or not the best word was within the top 1, 5 or 10 first words returned. The best word is the one with the smallest cosine distance to speech encodings. The code was run multiple times for each recall and the average was taken. This was done for both normal and synthetic data. On the left side of the above image in Table 2. we see that ADS generally slightly outperformes CDS. We also see that ADS generalizes better on CDS than the other way around. The same tendencies are also observed under synthetic data in Table 3. This tells us that the speech part doesn't have a large influence on the results and that it's mostly to do with the vocabulary differences. ADS generalizing better, was no surprise as it consists of a wider vocabulary and models trained on wider ranges of data tend to overfit less and generalize better on new unseen test data. On the right side in Figure 1. and 2. we see the first 10 epochs out of the 50 ran. CDS actually outperformed ADS in both natural speech data and synthetic data, more so in the latter. This makes sense as CDS has less variation between the utterances and because of that reaches best performances earlier. Because of the same small variation in the data it generalizes worse later on, especially when tested on ADS.

Model trained on CDS				
Testset	Med.r.	R@1	R@5	R@10
CDS	4.67	.28	.52	.61
ADS	52.50	.08	.19	.26
Combined	30.67	.15	.30	.37
Model trained on ADS				
Testset	Med.r.	R@1	R@5	R@10
CDS	37.83	.10	.24	.33
ADS	5.00	.29	.51	.61
Combined	20.83	.17	.32	.40

Table 2: Test performance of models trained on natural speech

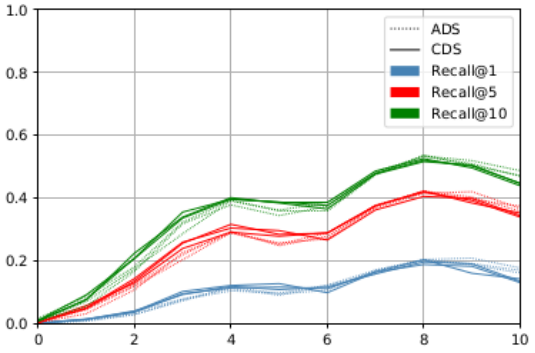

Figure 1: Validation performance in early training on natural speech

Model trained on synthetic CDS				
Testset	Med.r.	R@1	R@5	R@10
CDS	1.00	.82	.96	.99
ADS	1.00	.59	.79	.86
Combined	1.00	.68	.85	.90
Model trained on synthetic ADS				
Testset	Med.r.	R@1	R@5	R@10
CDS	1.00	.70	.89	.95
ADS	1.00	.84	.94	.97
Combined	1.00	.74	.89	.93

Table 3: Test performance of models trained on synthetic speech

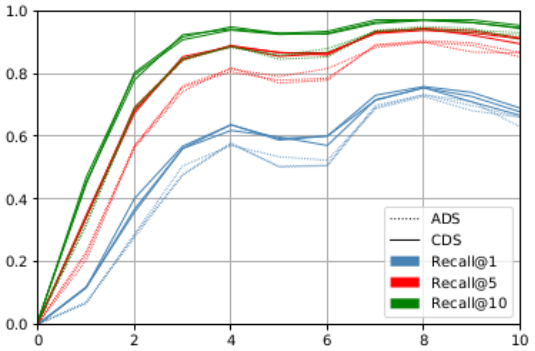

Figure 2: Validation performance in early training on synthetic speech

3. Shoulders of giants

When looking for similar papers there was actually a smaller amount than we expected to find. There was also some controversy as some papers found contrary results to this one. Kirchhoff and Schimmel (2005) showed that models trained on CDS work better on ADS than the other way around. Batchelder, 2002; Daland and Pierrehumbert, 2011 reported higher segmentability for CDS while Cristia et al. (2019) had mixed results. Data Science wise the paper they followed for the model was by Merkx et al. (2019) dealing in CNN with memory components that learned using images. The paper being reviewed went a step further (in our opinion) and used speech instead of images which is much easier to collect which means less resources are required.

4. Impact

The paper has been cited once, meaning it hasn't had much impact so far (by a paper trying to explain the mind of a child cartoon character in the popular cartoon "Family guy"). This is probably because nothing new was really discovered. All the results were fairly logical and the paper doesn't prove it well enough to be important in our opinion. Only recall was used and no accuracy or F1 score. The only thing we

think that could provoke more research was the use of direct speech embeddings when teaching models instead of turning speech into text first. This makes collecting data far easier for languages that aren't really common online, as it's much easier to record speech than to collect textual information. We're personally disappointed that the authors didn't try to do some more with the results, for example; training the model such that it goes through 10 to 20 epochs on CDS then switching to ADS trying to see whether it would learn faster in the beginning and generalize better later on. This way the best of both types of data would be used, mirroring the language learning capabilities of a child to adult. In addition, the paper doesn't mention it, but they used only speech where english was spoken. We feel there should have been more discussion about this. It would be interesting to see whether other languages give similar results. For example a language such as Japanese has special words that are more simple versions of complex ones which children use. Croatian has a special type of nouns that makes them sound "softer and smaller", thus more pleasing to children.

5. Reproducibility

At the time of presenting the paper, the code hadn't worked properly and we couldn't reproduce the results. Later on our professor had an email exchange with the author of the paper and the code was fixed and rerun. Among other issues there was a problem with data preprocessing. Namely spaces between words were changed with underscores when concatenating instead of the opposite. Fixing the code resulted in slightly lower values for everything but only by a little (between 0.01 and 0.02). Since all values were changed by more or less the same amount, this made no difference to the overall results. No appendix or supplementary data was given to give additional information about the research.

6. Summary of the received queries and of the in-class discussion

First use CDS and then ADS?
The number of words per second is twice as high in ADS as it is in CDS, does this effect learning?
Context of recorded speech (free play vs. interview)
Why do the authors only report recall, and not precision?
Only one algorithm was used (while in other studies multiple algorithms are used)
How can the results be generalized to other models for learning and detecting speech?
Are there implications of these results for other contexts outside of the context of this specific machine learning setup?

Do you think it is reasonable to use artificial language models (SBERT) to simulate human language learning?
Do you think that the use of semantic sentence embeddings is a valid replacement for visual data?

The class discussion actually took up most of the presentation time. Most people concluded that this paper shouldn't have been published for multiple reasons; some mentioned earlier. The first one being it's simply not good enough. It's a basic paper that would probably get a good grade if written by a college student but more is expected of a paper to be published in a high quality journal or conference. The paper has some new ideas, however, it doesn't do enough. It doesn't really bring new knowledge although it does contradict some earlier studies. A lot of students also complained about how little was done in terms of the used algorithms. Only one was used and only recall was shown (no F1 score or accuracy). I assume this lowered people's opinion of it the most as we can't be certain even in the results that have been shown. Then there's neglecting other languages. The authors don't address the issue saying that their study proves what it does only in the english language but you get the feeling it works for all languages.
Among other popular topics we discussed how good SBERT is for this given problem at all and why wasn't anything else was used to give more strength to their results.