Latent Feature Generation with Adversarial Learning for Aphasia Classification

Anna Vechkaeva, Günter Neumann
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{anna.vechkaeva, guenter.neumann}@dfki.de

Abstract

Aphasia is a language disorder resulting from brain damage, and can be categorised into types according to the symptoms. Automatic aphasia classification would allow for quick preliminary assessment of the patients' language disorder. A supervised approach to automatic aphasia classification would require substantial amount of training data, however, aphasia data is sparse. In this work, we attempt to use data generation, namely Generative Adversarial Networks (GANs), to deal with data sparsity. The latent feature generation approach is used to deal with the text generation non-differentiability problem, which is an issue for GANs. The approach using artificially generated data to augment training set was tested. We conclude through running a series of experiments that it has potential to improve aphasia classification in the context of low resource data, provided that the available data is enough for the generative model to properly learn the distribution.

Keywords: Adversarial Learning, Feature Generation, GANs, Aphasia

1. Introduction

Aphasia is a language disorder resulting from brain damage, such as stroke, physical damage, or degenerative dementias. Depending on the brain region, which was damaged, and the severity of the damage, it can manifest with various symptoms, which differ from patient to patient. Aphasia can be categorised into different types, which require different kinds of therapy. Therefore, automatic aphasia type classification could be beneficial for aphasia patients as well as speech therapists, as it would allow for quick preliminary assessment of patients' language disorder, and consequently faster therapy selection.

Although Allen et al. (2012) provides evidences that there exists effective therapy for chronic aphasia, multiple studies suggest that recovery after stroke (Kinsella and Ford, 1980; Skilbeck et al., 1983; Demeurisse et al., 1980), as well as after other brain damage (Jennett and Bond, 1975; Bond and Brooks, 1976) mostly happens in the first several months after the incident leading to the brain damage, and very little, if any, progress happens after one year (Hanson et al., 1989). Providing intensive therapy as soon as possible is crucial for rehabilitation and lack of it can compromise the outcome of the patients' recovery (Bhogal et al., 2003).

In this work, we attempt to classify the aphasia types in the context of low resource data. For this, deep neural networks (DNNs), as well as other machine learning algorithms were used. Using DNNs in this problem is challenging because they normally require a big amount of data to train successfully. One of the ways of dealing with data sparsity attempted in this work is generating synthetic data and using this data for training. Unlike (Chen et al., 2019) who generated structured data for patients' medical records, we focus on generating representations of unstructured textual data. We test if generating synthetic data can help improve the classification, focusing on the Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) framework due to their success in generative modeling. We adopt an adversarial feature learning approach (Ganin et al., 2016), which does not require generating actual textual data. Unlike GANs, which aim to generate realistic data, the adversarial feature learning approach generates the hidden representation of the data. This approach is suitable for non-generative tasks, such as classification, and alleviates the need of generating actual textual data, which notably has limitations with GANs.

In this work, we attempt to develop a model which, given a participant's speech transcript, predicts an aphasia type label. For this, we will first classify each individual phrase produced by the participant and use these labels to predict the participant's overall score. As the aphasia data is sparse we will also use a generative model to augment the training set. For this we will generate utterance level hidden representations, which will later be used for training utterance level classification model. The main contributions of this work are: (1) testing an approach of using GANs in the context of sparse data for aphasia classification, and (2) developing a latent feature generation approach to solve the GANs issues when dealing with text.

2. Background

2.1. Aphasia Classification

Aphasia was first described by French neuroanatomist Paul Broca (Broca, 1861) and since then, multiple aphasia classifications have been suggested. One of the ways to classify the patients into groups is using one of the standard protocols, for example, Western Aphasia Battery (WAB) (Kertesz, 2007). It distinguishes between the following aphasia types: Broca's, Wernicke's, anomic, conduction, transcortical motor, transcortical sensory, and global. These aphasia types differ by symptoms and severity. This aphasia classification scheme, as well as other ones, has been criticised, because often, patients' symptoms cannot be fit into one type and there exists overlap between the classes (Caramazza, 1984; Swindell et al., 1984). Nevertheless, WAB

provides a way to categorize patients according to their most prominent symptoms. Moreover there are datasets labeled using WAB scheme, which is important for studies using methods requiring substantial training data.

The speech of the people suffering from aphasia of different types and severity has distinctive characteristics, which can help to automatically analyze aphasia. There are studies which describe different features of aphasia speech in comparison to healthy speakers, as well as features specific to different aphasia types. These features can be acoustic (Damasio, 1992; Leung et al., 2017), grammatical (Kolk, 1998), discourse (Ulatowska et al., 1981) and semantic (Jefferies and Lambon Ralph, 2006). The presence of these features in aphasic speech suggests that it is possible to use them to automatically categorize aphasia.

Aphasia classification task is a problem which has been approached by researchers in the past. Järvelin and Juhola (2011) provide a system for distinguishing speech of people with aphasia from healthy controls' speech, and compare different machine learning techniques for identifying aphasia speakers. There are a number of studies, where authors attempt to distinguish different types of aphasia from each other, using groups of features. For example, Yourganov et al. (2015) attempt to predict types of aphasia based on fMRI brain images of the patients. There are studies, which assess aphasia, based on features extracted from other language production modalities like writing (Basso et al., 1978), sign language (Marshall et al., 2004), and comprehension (Mesulam et al., 2015; Purdy et al., 2019).

In order to analyze impaired speech, authors use two different kinds of features: acoustic and textual. For example, Qin et al. (2018) propose a system for assessing aphasia speech using textual features. The aphasia severity is predicted based on syllable level vectors, acquired from text produced by automatic speech recognition system, given recording of aphasia speech. Fraser et al. (2014) extract features from aphasia speech transcripts and use them to classify primary progressive (slow impairment of language caused by neurodegenerative disease) aphasia types. Themistocleous et al. (2018) identify mild cognitive impairment from speech using acoustic features. They predict if the patient has cognitive impairment based on features such as vowel formants, fundamental frequency and vowel duration. In Little et al. (2009) and Meilán et al. (2014) acoustic features were proven useful for detecting Parkinson's and Alzheimer's disease respectively from the patients' speech. Also, there are studies that provide evidence that combining acoustic and textual information helps to identify Alzheimer's disease (Fraser et al., 2016) and mild cognitive disease (Themistocleous et al., 2018). Although language impairments in case of dementias, can differ in their nature from ones in aphasia due to brain damage, similar approaches can be used to identify and assess them.

2.2. Synthetic Data Generation for Classification

The generative models are widely used to tackle the data sparsity problem in various fields and there is work on synthetic data generation for improving text analysis. For example, Maqsud (2015) tests different text generation methods for augmenting the available training data with synthetic samples for sentiment analysis of text. In this work, methods such as Latent Dirichlet Allocation (LDA), Markov Chain (MC), and Hidden Markov Model (HMM) are tested and the authors conclude that the models can generate the data with the features belonging to each class. In computer vision, synthetic data generation is also used to augment the sparse training data. The Generative Adversarial Networks (GANs) are used to improve different classification tasks, as GANs are able to generate realistic images. Frid-Adar et al. (2018) use GANs for generating the additional image data for improving liver lesion classification. In the paper, the authors train a separate generative model for each of the classes and then use the models to generate the data for the respective classes. A significant improvement of the classification after adding the generated data to the training set is reported. GANs have also been used to generate additional data for text classification. Guan et al. (2018) use conditional GANs to generate electronic medical records.

2.3. Generative Adversarial Networks

GANs were proposed by Goodfellow et al. (2014) and showed great success in image generation, becoming very popular. Given training data, the model learns the distribution of the data and produces data instances which belong to this distribution. GAN framework is derived from a game theoretic formulation, where each player can be seen as an adversary. GANs consist of two models: Generator and Discriminator. Given a noise (from normal distribution) as an input, generator's goal is to produce data samples which look like they belong to the same distribution as real data. When fed with the real data samples and samples produced by generator, discriminator's goal is to be able to distinguish between real and fake data. The discriminator's loss is then propogated back to generator so that it can improve and generate more realistic synthetic samples. The conditional GANs (cGAN), which were proposed by Mirza and Osindero (2014) are a type of GAN, that can be conditioned on some extra information. It learns to not only produce datapoints which look realistic but also conditions the produced datapoints with additional class information.

GAN models are known to have the training stability issues, meaning that the model does not converge. Other problems which may occur when using GANs are a mode collapse problem and the vanishing gradient problem (Goodfellow, 2016). A number of improved training techniques were proposed (Dziugaite et al., 2015; Huszár, 2015; Li et al., 2015; Salimans et al., 2016; Nguyen et al., 2017; Nowozin et al., 2016; Zhao et al., 2016) since the original introduction of GANs (Goodfellow et al., 2014). The most stable and robust version of GANs, called Wasserstein GAN (WGAN), was proposed by Arjovsky et al. (2017). Wasserstein GAN uses a different loss function for the discriminator, which is called critic in this setting. Instead of classifying the generated samples as real or fake, the critic tries to predict how close the produced samples are to the real distribution. Arjovsky et al. (2017) concluded that when using Wasserstein distance, problems such as mode collapse and vanishing gradient did not appear and the training was more stable.

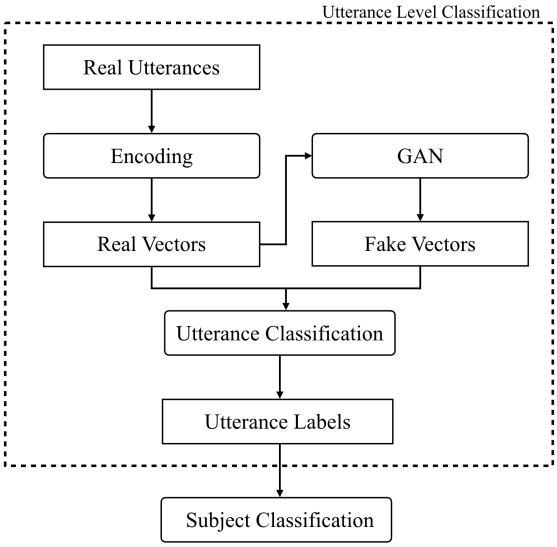

Figure 1: Classification Model with data generation

GANs have one design limitation on the Generator, that it cannot have discrete outputs. This makes them incompatible directly for NLP. By construction, the generative network has to be fully differentiable. Consequently, GAN framework prohibits generator from having discrete outputs. Also, it is not trivial to assign discriminator probabilities to the sequences which are not completely generated (Goodfellow, 2016). The Adversarial Feature Learning (AFL) approach is similar to the GANs approach. While in GANs the adversary aims to determine if the outputs are real or generated, in the AFL, the adversary is created over hidden features. This approach is well suited for non-generative tasks, like a sentence classification task, where the objective is not to classify sentences as real and fake. This approach allows the model to deal with continuous data, which is easier than dealing with the discrete outputs.

3. Method

3.1. Overview

For all experiments, the general approach taken in this work of classifying aphasia types is as follows: first, the utterances produced by a subject (or a person) are classified as one of the aphasia types, and after this, based on the utterance level classification, a subject is classified as healthy or having one of the aphasia types. This pipeline includes two classification models: utterance level classifier and subject level classifier. Different versions of both models were also tested in this work.

Figure 1 demonstrates the process of training the model involving artificial data generation. After encoding the real data into vectors, the conditional GAN model is trained using these vectors and corresponding labels. Then, using the trained GAN, the fake data vectors belonging to a specified class given the class label are generated. Following this, an utterance level classifier, trained on both real and generated data, is used to predict the utterance level labels on the test data, and the subject level classification is run, given the labels produced on the previous step, as an input, to predict the participant level aphasia type labels. The reason for doing utterance level classification first and then subject level classification, instead of doing the subject level classification directly, is that Neural Network approaches normally require a sufficient amount of training data. While the number of subjects in the dataset used in this work is small, the number of utterance level datapoints is much bigger.

3.2. Data

AphasiaBank (MacWhinney et al., 2011) is one of the few publicly available datasets in the aphasia domain. It contains recordings of people suffering from aphasia as well as transcripts of their speech which also provide information about a patient including aphasia type. AphasiaBank also includes interviews with healthy participants recorded following a similar procedure. Table 1 shows an example of a transcript of speech belonging to a patient with anomic aphasia.

Anomic Aphasia
1 INV: how do you think your speech is these days ?
2 PAR: uh it's ... it's good but it's very slow .
3 INV: do you remember when you had your stroke ?
4 PAR: um it's two years ago .
5 PAR: and when I when I when I had the stroke I couldn't say a word for a year and a half .

Table 1: Example of utterances produced by a patient with Anomic Aphasia (source: AphasiaBank)

The aphasia type labels and aphasia severity scores provided in AphasiaBank are obtained using Western Aphasia Battery (WAB) (Risser and Spreen, 1985). WAB is an instrument for evaluating clinical aspects of language function for individuals with neurological disorders resulting from stroke, brain injury or dementia. It helps to identify presence, severity and type of aphasia and measures linguistic (speech, fluency, auditory comprehension, reading and writing) and non-linguistic performance of individuals.

The dataset, which was constructed in this work, consists of utterance level data-points, which are transcriptions of the phrases produced by a subject in response to the interviewer's question. The utterance level datapoints are grouped into subject level datapoints and each of them consists of transcribed utterances produced by one subject. This is done so that the subjects as well as utterance level classification can be performed. AphasiaBank does not provide utterance level aphasia type labels, so the labels for the utterances are assigned based on the aphasia type of the participant who produced the utterance.

Not all utterances, produced by patients suffering from aphasia, contain signs indicating the aphasia type, some of them are completely correct, as shown in the example in Table 1, where utterance 2 is grammatically correct. AphasiaBank provides aphasia type labels only on a subject level, but not on the utterance level. The fact that not all of the

	Utterance level		Subject level	
	Train	Test	Train	Test
Broca's	2682	1029	66	30
Anomic	10767	2814	108	30
Conduction	4141	3859	34	30
Control	29969	4073	217	30
Total	47559	11775	425	120

Table 2: Number of utterance and subject level datapoints for training and test set

utterances by aphasia patients are aphasic provides a challenge for constructing an utterance level dataset, as aphasia type labels for each utterance cannot be confidently determined. This introduces a certain level of noise to both training and test datasets, as there are a number of non-aphasic utterances marked as aphasic ones. This provides challenges for both model training and the evaluation of results.

The dataset used for training and evaluating our models included the classes, which contained at least 60 patients, so that at least 30 subjects could be used in the test set. The final dataset contained the following classes: Broca's, Anomic, Conduction and Control. The number of both utterance and subject level datapoints belonging to each class are represented in Table 2.

3.3. Hidden Text Representation

In this work, 300 dimensional word vectors pretrained by Google using word2vec (Mikolov et al., 2013) were used for generating hidden text representation [1]. Each utterance from the dataset is represented as a two dimensional matrix constructed from the vectors of each word. In order to make all of the utterance representations have the same dimensions, we add 300 dimensional vectors of zeros to the utterance representations until all representations have the same dimensionality.

3.4. Models

3.4.1. Utterance Classification

A Convolutional Neural Network (CNN) (LeCun et al., 1989), which takes a two dimensional vector of an utterance as input and produces probabilities of the utterance belonging to each of the considered classes, was used for utterance classification in all the experiments. The architecture of the model is the same for all the experiments and contains three layers, with the first layer being a linear layer, which flattens the input. This is done so that the vertical relations between the values in word vectors are not taken into account, as, intuitively, unlike in the numerical image representations, there should be no vertical correlation between the individual values of the word vectors. The next layer is a layer with the ReLU activation function.

3.4.2. Subject Classification

The aim of a subject level classification model is to predict a participant's aphasia label, given utterance level labels predicted for this participant by the utterance level classifier.

[1] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

In this work, we test two approaches: supervised and unsupervised. The advantage of the unsupervised models is that they do not require additional data to train, therefore the whole training dataset can be used to train the generation and utterance level classification models. On the other hand, we expect that the supervised models will show better results than the unsupervised models, as they will observe the actual distribution of the utterance level labels over the training data. However, as a part of the data will need to be held out during the steps presiding the subject level training, the quality of the generated data can drop.

In an unsupervised approach, a fairly simple model was used. After all the utterances were classified by the utterance level model, the subject was assigned the aphasia type which was most present among the subject's utterances, according to the classifier. For example, if a subject had 20 utterances classified as Broca's aphasia, 30 utterances - as Anomic aphasia, 10 utterances - as Conduction aphasia, and 5 sentences - as non-aphasic sentences, the patient was classified as having Anomic aphasia. This algorithm will be further referred to as the Max Class model. Different variations of the described model were also tested. As patients suffering from aphasia are still able to produce non-aphasic sentences, it makes sense to reduce the impact of the non-aphasic utterance on the patient level classification. So, instead of taking into account the entirety of non-aphasic sentences, this number is reduced by dividing it by a range of integers from 2 to 7.

The number of supervised machine learning algorithms were also tested to predict an aphasia type on the subject level, given the utterance level labels predicted by the utterance level classifier. Given a number of utterances for which an utterance level classifier predicted each type of aphasia, it predicted a patient level aphasia type. The algorithms used are Naïve Bayes (NB), Multinomial Naïve Bayes (MNB), Random Forest (RF), Decision Trees (DT), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM).

3.4.3. Generation Model

In this work, we aim to generate hidden features for synthetic data without generating the actual data for text classification using cGANs. We use the static text representations to train the generative model, to produce a hidden representation of the data from the different classes. The model does not take into account other subject properties, like age or gender. That way, the model aims to generate a vector representation of an utterance belonging to a given class without generating actual textual data. The fact that we create the adversary over the hidden features makes this approach similar to the AFS.

For data generation, two types of GANs were tried. The first one is a simple conditional GAN. Both generator and discriminator of this model are CNNs. Binary crossentropy is used as a loss function and the GAN is conditioned on aphasia type and produces vector representation of utterance level datapoints, given a class. The models trained for different amount of epochs were tested. Conditional WGAN, where both generator and critic are CNNs and Wasserstein loss is used, is also tested. It is also condi-

tioned on aphasia type labels. The models trained for the different amount of epochs are also tested. In this work, the Keras implementation of GANs[2] for cGAN was and for WGAN was used.

3.5. Experiments

3.5.1. Baseline

The baseline system uses only real data from AphasiaBank as training data, and no data generation happens at this step. A CNN model trained on the whole training set containing real utterance level datapoints was used to predict an aphasia class for an utterance given its vector representation. For the subject level classification, the Max Class model without any additional alternations was used.

3.5.2. GAN models comparison

The different GAN models that were trained for a different number of epochs were compared, as it is difficult to tell if the GAN converged based only on generator and discriminator losses and sometimes GANs mode collapse can occur, which means that the generator starts producing very similar outputs to trick the discriminator. The conditional GAN trained for 20000 epochs and 5000 epochs and Wasserstein conditional GAN trained for 500 epochs, 1000 epochs, and 2000 epochs models were compared.

In order to assess how good the data produced by each of the tested models is, we investigate how this data can help with the aphasia classification problem. The produced data was used to train an utterance level CNN classifier which was later used to predict aphasia classes for utterances belonging to patients from the test set. After this, these predictions were used by the Max Class model to assign an aphasia type to each patient in the training set.

For each of the models, two different experiments were performed. The first experiment aims to investigate how good the performance of the classification model trained only on generated data is. Each of the GAN models described above was trained on the whole training set to produce utterance level datapoints belonging to a given aphasia class. The generator model produced by each of the models was then used to generate synthetic training data, producing 40000 datapoints (10000 datapoints for each class) by each of the models. Then, an utterance level CNN classifier was trained using the generated data as a training set and max class model was used to predict patient level classes.

The purpose of the second experiment was to assess how combining generated data with the real data can help improve the aphasia type classification. For this experiment, the synthetic data was generated for the aphasia types which have fewer utterance level datapoints in the AphasiaBank, so that each class has the same number of datapoints, resulting in 27 287 generated datapoints for Broca's aphasia, 19 202 - for Anomic aphasia, and 25 828 - for Conduction aphasia. The control group contains the biggest amount of datapoints, so no data was generated for this class. After the data generation step, all the produced datapoints were added to the original training set and each class ended up having 29 969 utterance level datapoints. After this, the CNN utterance level classifier was trained on the combined dataset and the Max Class model was used to predict the aphasia type of the subject.

[2]https://github.com/eriklindernoren/Keras-GAN

	Utterance level		Subject level	
	Train1	Train2	Train1	Train2
Broca's	5433	5406	33	33
Anomic	1346	1409	54	54
Conduction	2140	2264	17	17
Control	15115	15309	108	109
Total	24034	24388	212	213

Table 3: Number of utterance and subject level datapoints for two training tests

3.5.3. Max Class Model Experiments

As not all of the utterances produced by aphasia patients show signs of aphasia, reducing the impact of the control class on the subject level classification could help to improve the classification. In order to reduce the impact of the control class, we use the number of the control class predictions divided by some number instead of the full number. For example, if the majority of the utterances produced by a subject are classified as non-aphasic, but some other aphasia type class is very present amongst the utterances, the subject might still have aphasia.

The aim of this experiment is to determine by how much the impact of the control class should be reduced. To determine by how much the number of predicted control class should be divided, a range of numbers from 1 to 7 are tested. We assume that initially the classification accuracy will increase as the number becomes bigger, but will start dropping after a point when the impact of the control class will become too small. We also compare the performance of each Max Class model trained on the real data with the one trained on the combination of real and generated data.

3.5.4. Supervised Subject Classification Methods Experiments

In addition to Max Class model a number of machine learning (ML) algorithms were tested for the subject level classification: NB, MNB, RF, DT, KNN, and SVM. Given the labels produced by an utterance level classifier, the classifier should predict the aphasia type of the subject. The Scikit Learn (Pedregosa et al., 2011) implementation of the models listed above was used.

This approach required splitting training data into two parts, as the algorithms used need data to be trained on as well as the utterance level classifier. The utterance level classifier and the subject level ML models need to be trained on different data, because if both of the models will be trained on the same data, the CNN classifier will have to make predictions, which will later be used for training by the ML algorithms for the samples it observed during the training. Our concern is that, given that the data is noisy, the prediction quality for this data will be too different from the predictions made for unseen data. Similarly, the GAN model should not be trained on the data, which later will be used for the ML models training, because if the generative model produces data similar to the data which will be later used for the subject level model training, the model will still

Model	Utt. level	Subj. level
Baseline	0.45	0.44
GAN 5000 epochs	0.46	0.39
GAN 20000 epochs	0.46	0.43
W-GAN 500 epochs	0.46	0.42
W-GAN 1000 epochs	0.45	0.48
W-GAN 2000 epochs	0.46	0.45

Table 4: Classification accuracy for GAN models trained for different amount of epochs (trained on both real and generated data)

be indirectly trained on this data through the GAN, and the predictions on this data will be different from the predictions on the unseen data. Therefore, the data was divided subject-wise into two equal parts, so that each part contains the same number of patients per aphasia type class. The number of utterance and subject level datapoints for each class is presented in Table 3. The first half of the data contains 212 subjects and 24 034 utterances, while the second half of the data contains 213 subjects and 24 388 utterances. To compare the performance of the subject level classification models with the approach used before, we also run the Max Class models on the newly divided data. We used only the first part of the training data to train CNN classifier, while the second part of the data was not used, as Max Class models do not need training.

4. Results

4.1. Evaluation

To assess the performance of the models, accuracy and F1 score are used. On the utterance level only the accuracy is reported. The performance of the models, evaluated on the utterance level test set, does not really reflect the quality of the models. The reason for this is that the test set contains noisy data, because the subject level labels are used to assign labels to utterances when constructing the test set.
Because the dataset has gold-standard labels on the subject level, unlike the utterance level evaluation, the subject level evaluation reflects the quality of the system. For the subject level evaluation classification accuracy is reported, and for the final comparison of the models performance, F1 scores are reported. These metrics are reported for each class as well as for the whole test dataset.

4.1.1. GANs Comparison

Table 4 shows the utterance and patient level accuracy for the classification model which used data generated by different GAN models in addition to the real data. The baseline is the model trained only on real data. The Wasserstein GAN trained for 1000 epochs demonstrates the best results on the patient level, showing 4% accuracy improvement over baseline. The Wasserstein GAN model trained for 500 epochs and simple GAN models trained for 5000 and 20000 epochs perform worse than baseline. The Wasserstein GAN trained for 2000 and the simple GAN trained for 20000 epochs demonstrates 1% improvement over baseline.
The utterance and subject level accuracy for the classification model trained using only the data generated by different GAN models is shown in Table 5. The baseline is

Model	Utt. level	Subj. level
Baseline	0.45	0.44
Random Classifier	0.25	0.25
GAN 5000 epochs	0.24	0.24
GAN 20000 epochs	0.33	0.29
W-GAN 500 epochs	0.30	0.31
W-GAN 1000 epochs	0.35	0.31
W-GAN 2000 epochs	0.35	0.25

Table 5: Classification accuracy for GAN models trained for different amount of epochs (trained only on generated data)

the classification model trained only on real training data from AphasiaBank. The results for the random classification are also reported. The results show that Wasserstein GAN trained for 1000 epochs and for 500 epochs demonstrate the best improvement over random classifier, however they do not beat the baseline. The simple GAN trained for 20000 epochs also demonstrated small improvement over random classifier. Simple GAN trained for 5000 epochs and Wasserstein GAN trained for 2000 epochs show performance similar to random classifier. As Wasserstein GAN trained for 1000 epochs demonstrates the best performance in the both experiments, it is used in all the following experiments to generate synthetic data.

4.1.2. Max Class Experiments

	Real	Real + Generated
MC-1	0.35	0.41
MC-2	0.38	0.48
MC-3	0.42	0.49
MC-4	0.43	**0.53**
MC-5	0.45	0.49
MC-6	0.46	0.47
MC-7	0.42	0.44

Table 6: F1 for the models trained on the real and combined data using the unsupervised Max Class models for subject level classification

Table 6 represents the results of the different Max Class models with the utterance classifier trained on real data and on combination of real and generated data. The table presents the F1 score for each of the Max Class models. The results show that the best performing model is the model trained on the combination of real and generated data on the utterance level and divides the number of the predicted non-aphasic utterance by 4. Also, the table shows that reducing the impact of the non-aphasic class helps to improve the classification.
Table 7 shows the results of the models using Max Class on the subject level for each aphasia class. The results show that none of the tested models managed to classify conduction aphasia and the models do not ever predict the conduction aphasia class. Reducing the impact of the non-aphasia class improved classification of control group and Anomic aphasia group, and did not influence classification

Model	Anom.	Broc.	Cond.	Contr.
MC-1 Real	0.41	0.34	0.00	0.64
MC-1 Comb.	0.38	0.61	0.00	0.65
MC-2 Real	0.41	0.33	0.00	0.76
MC-2 Comb.	0.46	0.60	0.00	0.87
MC-3 Real	0.44	0.33	0.00	0.88
MC-3 Comb.	0.46	0.60	0.00	0.91
MC-4 Real	0.45	0.33	0.00	0.91
MC-4 Comb	0.54	0.60	0.00	0.98
MC-5 Real	0.51	0.33	0.00	0.97
MC-5 Comb	0.51	0.60	0.00	0.87
MC-6 Real	0.51	0.33	0.00	0.98
MC-6 Comb	0.49	0.60	0.00	0.78
MC-7 Real	0.49	0.33	0.00	0.87
MC-7 Comb	0.48	0.60	0.00	0.71

Table 7: Individual F1 per class for the models trained on real and combined data using Max Class models for the subject level classification

of Broca's and Conduction aphasia. Adding generated data improves classification of Broca's aphasia. For Anomic aphasia, generating data improved F1 score.

4.1.3. Supervised Methods Experiments

	Real	Real + Generated
Max Class	0.34	0.32
Naive Bayes	0.42	0.46
Multinomial NB	0.52	0.58
Decision Trees	0.55	0.52
Random Forest	**0.61**	0.53
KNN	0.56	0.56
SVM	**0.61**	0.58

Table 8: F1 for the models trained on the real and combined data using the supervised models for subject level classification

Table 8 shows that, although adding the generated data to the training set for utterance level classification and synthetic data generation helps when the Naive Bayes classification is used for the subject level classification improving the results of classification from F1=0.42 to F1=0.46 and F1=52 to F1=58 for Gaussian Naive Bayes and Multinomial Naive Bayes classifiers respectively, the data generation did not improve the results for other subject level calssification methods. Out of all the methods used, the SVM and RF classification on the subject level without data generation showed the best results (F1= 0.61). For this methods, generating the additional data did not help to improve the classification. Also, unlike the previous experiments where both generator and utterance level classifier were trained on the whole dataset, the results for the Max Class utterance level classification did not improve when the generated data was added to the training set. For the Max Class, DT, RF and SVM subject level classification models adding the generated data made the results worse. And for the KNN classifier the F1 score stayed the same when the generated data was included in the training set.

Model	Anom.	Broc.	Cond.	Contr.
MaxClass Real	0.42	0.19	0.00	0.75
MaxClass Comb.	0.38	0.29	0.00	0.59
NB Real	0.35	0.63	0.06	0.64
NB Comb	0.31	0.62	0.23	0.67
Mult. NB Real	0.46	0.61	0.11	0.91
Mult. NB Comb.	0.42	0.63	0.39	0.88
DT Real	0.46	0.67	0.16	0.91
DT Comb.	0.39	0.72	0.12	0.85
RF Real	0.47	0.77	0.31	0.89
RF Comb.	0.43	0.73	0.12	0.87
KNN Real	0.46	0.70	0.18	0.90
KNN Comb	0.49	0.77	0.12	0.86
SVM Real	0.53	0.75	0.18	0.97
SVM Comb.	0.47	0.74	0.18	0.94

Table 9: Individual F1 per class for the models trained on real and combined data using the suppervised models for the subject level classification

Table 9 shows the results for the models trained on the real data and combination of the real and generated data for each aphasia class. It shows that unlike the Max Class methods, the supervised methods manage to sometimes predict the conduction aphasia class. However, the F1 score for this class still performed the worst out of all the classes.

5. Discussion

The Wasserstein GAN trained for 1000 epochs produced the best results. Wasserstein GAN trained for 500 epochs produced worse results because it probably did not converge, meaning that both discriminator and generator were not good enough to produce data resembling real data. Wasserstein GAN trained for 2000 also performed worse than the one trained for 1000 epochs. It likely means that the mode collapse problem occurred, meaning that the generator learned to produce output datapoints which were not diverse, but managed to trick the discriminator. The simple GAN trained for 5000 epochs performed the worst out of all the trained models. This model did not mange to converge, as empirically, it takes longer for the original GAN to converge due to possible oscillations in optimization, whereas WGAN has more stable training, leading to faster optimization. The simple GAN model trained for 20000 epochs performed better than the one trained for 5000 epochs. These results match our intuitions that Wasserstein GANs converge faster than simple GANs.

Supervised machine learning methods for the subject level classification outperformed unsupervised methods. Although when using the models with unsupervised subject classification, augmenting the training set with the generated samples improved the classification, the highest result for the Max Class model with the reduced impact of the non-aphasia class (F1 = 0.53) was outperformed by Multinomial NB, DT, RF, KNN, and SVM classification methods trained only on real data. The RF and SVM sowed the best result. Adding the generated data to the training set improved the performance of the model only for Multino-

mial NB (from F1=0.52 to F1=0.58) and NB (from F1=0.42 to 0.46) models. For the other models, including the Max Class model, the performance stayed the same (KNN) or dropped (Max Class, DT, RF, and SVM). Using the generated data did not help to beat the best performing model trained only on the real data. The reason for this may be that when using the supervised machine learning techniques on the subject level, the training data has to be split in two parts which leads to reducing the training set for generative model. It is possible, that with the reduced amount of training data the model did not manage to learn to generate samples diverse enough for helping the classification. The fact that data generation improved the performance of the simple Max Class model when trained on the whole training dataset, and failed to improve the performance of the same model when trained on the reduced dataset supports that explanation.

All tested Max Class models failed to classify Conduction aphasia and for the ML classifiers which managed to predict the Conduction aphasia class, the F1 score for the Conduction aphasia is lower than for the other classes. In AphasiaBank, Conduction aphasia has the least amount of patients. Possibly, the data was not diverse enough to classify this type of aphasia and generate good artificial data. Conduction aphasia almost always classified as Anomic aphasia. Anomic and Conduction aphasia are fairly similar in writing: both are characterised by fluent speech. In addition, the WAB aphasia severity scores for Conduction and Anomic aphasia patients are quite close, which means that these types of aphasia have similar level of severity. While patients with Anomic aphasia often use neologisms and frustration markers, patients with conduction aphasia often produce words incorrectly. In both these cases, the produced words will be treated as OOV words by the classifier and will not be accounted for.

Intuitively better classification on the utterance level should lead to the better classification on the subject level. However, this is not the case for the current experiments. The reason for this is that the test set we are evaluating the utterance level classification on is noisy, because of the aphasia patients producing non-aphasia utterances. Therefore the classification accuracy on the utterance level does not really reflect the real quality of the classification. So, there are cases when although the classification on the utterance level improves the classification on the subject level drops and other way round.

6. Related Work

Most of the works focused in the aphasia or mild cognitive disease classification tend to treat this problem as a binary classification problem. A lot of studies focus on the impaired and non-impaired speech classification (Järvelin and Juhola, 2011; Themistocleous et al., 2018; Little et al., 2009; Meilán et al., 2014). The others try to distinguish one type of language impairment from another, still treating the problem as binary classification (Fraser et al., 2014; Yourganov et al., 2015). Therefore, the results reported in these works cannot be directly compared to our results in the current setting. To the best of our knowledge, the classification of multiple aphasia types has not been attempted by researchers.

However, approaches, similar to the one taken in this work, were tested in different domains and these results can be indirectly compared to ours. For example, Guan et al. (2018) used cGANs to augment training data for automatic electronic medical records (EMR) classification into diagnosis types. The task in their work is similar to ours, because they also compare the the models trained only on real data with the models trained on the combination of real and generated data. The dataset used contained 2216 EMR texts which were assigned one of the two diagnosis: pneumonia and lung cancer. For data generation, the authors use a model called Medical Text GAN (mtGAN) which generated text samples using reinforcement learning to solve the text non-differentiability problem. Guan et al. (2018) report that after adding the generated data to the real training set the classification accuracy improved from 0.7500 to 0.7635 (0.0135 improvement).

Although from the high level perspective our approach is similar to the one used by Guan et al. (2018), it differs in details. First their data contains texts written by doctors about patients, while we focus on the speech produced by the patients. Second, the different strategies are used due to the structural differences of the data; while we use two level classification, Guan et al. (2018) classify the EMRs directly. Finally, our approach to data generation is different, as we generated the data on the hidden representation level, while Guan et al. (2018) generated the textual samples. In the case of the unsupervised subject classification, our results demonstrate bigger improvement when using generated data in combination with real data. The best system using Max Class system and only real data demonstrate the accuracy of 0.46, while adding the generated data brings the accuracy to 0.53 (0.07 improvement). The bigger improvement in the aphasia classification case could be caused by the difference in the approaches as well as by the difference in the datasets.

7. Conclusion

The method of using the same text representations for both generation and classification tasks was proposed. By encoding the text into vectors from the beginning and then generating and classifying vector representations, we avoid the problem of text being discrete when using GANs. Also, this approach requires only encoding the text, but no decoding is needed.

The results show that using hidden feature generation with GANs for improving text classification is useful in certain cases, and generating additional synthetic data and combining it with the real data for training improves the classification results. However, it has certain limitations, namely, the generation model still needs sufficient amount of data to be able to produce useful output.

8. Acknowledgements

This work was partially funded by the European Union's Horizon 2020 grant agreement No. 777107 (Precise4Q) and by the BMBF project DeepLee (01IW17001).

9. Bibliographical References

Allen, L., Mehta, S., Andrew McClure, J., and Teasell, R. (2012). Therapeutic interventions for aphasia initiated more than six months post stroke: a review of the evidence. *Topics in stroke rehabilitation*, 19(6):523–535.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. arxiv e-prints, page. *arXiv preprint arXiv:1701.07875*.

Basso, A., Taborelli, A., and Vignolo, L. (1978). Dissociated disorders of speaking and writing in aphasia. *Journal of Neurology, Neurosurgery & Psychiatry*, 41(6):556–563.

Bhogal, S. K., Teasell, R., Speechley, M., and Albert, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke-a Journal of Cerebral Circulation*, 34(4):987–991.

Bond, M. and Brooks, D. (1976). Understanding the process of recovery as a basis for the investigation of rehabilitation for the brain injured. *Scandinavian Journal of Rehabilitation Medicine*.

Broca, P. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6:330–57.

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and language*, 21(1):9–20.

Chen, J., Chun, D., Patel, M., Chiang, E., and James, J. (2019). The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC medical informatics and decision making*, 19(1):44.

Damasio, A. R. (1992). Aphasia. *New England Journal of Medicine*, 326(8):531–539.

Demeurisse, G., Demol, O., Derouck, M., De Beuckelaer, R., Coekaerts, M., and Capon, A. (1980). Quantitative study of the rate of recovery from aphasia due to ischemic stroke. *Stroke*, 11(5):455–458.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, 55:43–60.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Guan, J., Li, R., Yu, S., and Zhang, X. (2018). Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380. IEEE.

Hanson, W. R., Metter, E. J., and Riege, W. H. (1989). The course of chronic aphasia. *Aphasiology*, 3(1):19–29.

Huszár, F. (2015). How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*.

Järvelin, A. and Juhola, M. (2011). Comparison of machine learning methods for classifying aphasic and non-aphasic speakers. *Computer methods and programs in biomedicine*, 104(3):349–357.

Jefferies, E. and Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain*, 129(8):2132–2147.

Jennett, B. and Bond, M. (1975). Assessment of outcome after severe brain damage: a practical scale. *The Lancet*, 305(7905):480–484.

Kertesz, A. (2007). *WAB-R: Western aphasia battery-revised*. PsychCorp.

Kinsella, G. and Ford, B. (1980). Acute recovery from patterns in stroke patients: neuropsychological factors. *The Medical Journal of Australia*, 2(12):663–666.

Kolk, H. (1998). Disorders of syntax in aphasia: Linguistic-descriptive and processing approaches. In *Handbook of neurolinguistics*, pages 249–260. Elsevier.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Leung, J. H., Purdy, S. C., Tippett, L. J., and Leão, S. H. (2017). Affective speech prosody perception and production in stroke patients with left-hemispheric damage and healthy controls. *Brain and language*, 166:19–28.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE transactions on bio-medical engineering*, 56(4):1015.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Maqsud, U. (2015). Synthetic text generation for sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 156–161.

Marshall, J., Atkinson, J., Smulovitch, E., Thacker, A., and Woll, B. (2004). Aphasia in a user of british sign language: Dissociation between sign and gesture. *Cognitive neuropsychology*, 21(5):537–554.

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Mesulam, M.-M., Thompson, C. K., Weintraub, S., and Rogalski, E. J. (2015). The wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. *Brain*, 138(8):2423–2437.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., and Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Purdy, M., Coppens, P., Madden, E. B., Mozeiko, J., Patterson, J., Wallace, S. E., and Freed, D. (2019). Reading comprehension treatment in aphasia: A systematic review. *Aphasiology*, 33(6):629–651.

Qin, Y., Lee, T., Feng, S., and Kong, A. P. H. (2018). Automatic speech assessment for people with aphasia using tdnn-blstm with multi-task learning. *Proc. Interspeech 2018*, pages 3418–3422.

Risser, A. H. and Spreen, O. (1985). The western aphasia battery. *Journal of clinical and experimental neuropsychology*, 7(4):463–470.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

Skilbeck, C. E., Wade, D. T., Hewer, R. L., and Wood, V. A. (1983). Recovery after stroke. *Journal of Neurology, Neurosurgery & Psychiatry*, 46(1):5–8.

Swindell, C. S., Holland, A. L., and Fromm, D. (1984). Classification of aphasia: Wab type versus clinical impression. In *Clinical Aphasiology: Proceedings of the Conference 1984*, pages 48–54. BRK Publishers.

Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2018). Identification of mild cognitive impairment from speech in swedish using deep sequential neural networks. *Frontiers in neurology*, 9.

Ulatowska, H. K., North, A. J., and Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and language*, 13(2):345–371.

Yourganov, G., Smith, K. G., Fridriksson, J., and Rorden, C. (2015). Predicting aphasia type from brain damage measured with structural mri. *Cortex*, 73:203–215.

Zhao, J., Mathieu, M., and LeCun, Y. (2016). Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*.