Cross-Language Aphasia Detection using Optimal Transport Domain Adaptation

Aparna Balagopalan   APARNA@WINTERLIGHTLABS.COM
WinterLight Labs

Jekaterina Novikova   JEKATERINA@WINTERLIGHTLABS.COM
WinterLight Labs

Matthew B. A. McDermott   MMD@MIT.EDU
Massachusetts Institute of Technology

Bret Nestor   BRETNESTOR@CS.TORONTO.EDU
University of Toronto, Vector Institute

Tristan Naumann   TRISTAN@MICROSOFT.COM
Microsoft Research

Marzyeh Ghassemi   MARZYEH@CS.TORONTO.EDU
University of Toronto, Vector Institute

Editors: Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

Multi-language speech datasets are scarce and often have small sample sizes in the medical domain. Robust transfer of linguistic features across languages could improve rates of early diagnosis and therapy for speakers of low-resource languages when detecting health conditions from speech. We utilize out-of-domain, unpaired, single-speaker, healthy speech data for training multiple Optimal Transport (OT) domain adaptation systems. We learn mappings from other languages to English and detect aphasia from linguistic characteristics of speech, and show that OT domain adaptation improves aphasia detection over unilingual baselines for French (6% increased F1) and Mandarin (5% increased F1). Further, we show that adding aphasic data to the domain adaptation system significantly increases performance for both French and Mandarin, increasing the F1 scores further (10% and 8% increase in F1 scores for French and Mandarin, respectively, over unilingual baselines).

1. Introduction

Aphasia is a form of language impairment that affects speech production and/or comprehension. It occurs due to brain injury, most commonly from a stroke, and affects up to 2 million people in the US alone NAA (2016). Evaluation of speech is an important part of diagnosing aphasia and identifying sub-types. Aphasic speech exhibits several common patterns; e.g., omitting short words ("a", "is"), using made-up words, etc. Prior work has shown that it is possible to detect aphasia with machine learning (ML) from patterns of linguistic features in spontaneous speech Fraser et al. (2014), but a vast majority of research is restricted to a single language Fraser et al. (2014); Le et al. (2017); Qin et al. (2018).

Cross-linguistic studies for aphasia detection and screening Bates et al. (1991); Soroli et al. (2012); Kristen et al. (2014) are important for translating developments made in resource-rich languages (e.g., English) to other languages. Cross-lingual aphasia detection is a hard task, mainly because there is little prior research indicating what features of language affected by impairment are transferable and due to the small size of datasets in the field (typically between 100-500 subjects).

Existing research on cross-language translation of text, word embeddings and audio has demonstrated the value in using large amounts of paired and unpaired data (with dataset size varying from 5k to 1.5M) by imposing constraints such as cycle-consistency or incorporating domain-knowledge Xu et al. (2018); Yang et al. (2018); Conneau et al. (2018); Jia et al. (2019); Weng and Szolovits (2018); Weng et al. (2019). These allow translation of representations across languages for a variety of tasks, even with *no aligned data* between languages, which is of interest for our prediction task.

In this work, we study cross-linguistic transfer of aphasia detection models trained on English speech from a multi-lingual dataset of healthy and aphasic speech, Aphasia-Bank MacWhinney et al. (2011). We featurize our speech via the proportions of 8 standard linguistic Part-of-Speech (POS) tags from speech transcripts of each language using the StanfordNLP library Qi et al. (2018), motivated by prior work on automatic aphasia detection with text features Fraser et al. (2014). We adapt features from different languages to English using a state-of-the-art technique for domain adaptation, Optimal Transport (OT), with a large *single-speaker, unpaired multilingual dataset* of TED talk transcripts Tiedemann (2012). OT is a method of domain adaptation where the cost of moving samples from source to target probability distributions is minimized (Sec. 4.1). In our case, probability distributions are the distributions of linguistic features in each language.

We benchmark performance of models Pedregosa et al. (2011); Flamary and Courty (2017) trained on English AphasiaBank and tested on French and Mandarin AphasiaBank using different levels of capacity, including support vector machines (SVM), Random Forests (RF), and fully-connected neural networks (NN). We compare unilingual models, which suffer due to the limited size of the French and Mandarin datasets, to two kinds of domain adaption: a baseline multi-lingual autoencoder, and OT based domain adaption. While a multi-lingual autoencoder works well for the similar languages of French and English, it performs poorly for Mandarin. In contrast, OT-based adaption systems perform well across both languages. We additionally study the impact of various settings on the OT-domain adaptation system, such as inclusion of paired data, addition of in-domain data and speech-accent, against both unilingual baselines and an autoencoder-based domain adaption model. We show that OT-domain adaptation with Earth Movers Distance (EMD) and entropic regularization achieves 10% and 8% increase in F1 scores over the unilingual baselines in French and Mandarin respectively. From our analysis, we identify that large datasets, which could be unpaired across languages, but also include some samples of aphasic speech are essential for higher detection rates.

Healthcare Relevance: Medical speech datasets for many languages are small and few in number. As a result, most of the prior work on computational methods for detecting signs of aphasia focuses on ML-models developed for a single language Fraser et al. (2014); Le et al. (2017); Qin et al. (2018), or cross-language feature patterns for a single feature Bates

et al. (1991). In this work, we study both feature transfer across similar and dissimilar languages when participant speech is represented by multiple linguistic features, as well as cross-linguistic transfer of ML-models for detecting aphasia from linguistic features for translating developments made in English to other languages.

Technical Sophistication: We benchmark the utility of unpaired speech datasets with OT to demonstrate a use of cross-language domain adaptation to account for sparsely labeled languages in a clinical speech task. We perform rigorous ablation studies to investigate the effect of paired data (inclusion of paired data does not increase F1-scores significantly), the effect of aphasic samples in domain adaptation (improves F1-scores significantly over unilingual baselines and domain adaptation with unpaired data, performs on the same or significantly higher level than multi-lingual baselines) and the effect of diversity of speech samples in terms of accents (more diverse data for OT is better). While we only compare OT-domain adaptation with a multi-lingual autoencoder, and do not compare against other, more recent techniques, such as adversarial domain adaptation Tzeng et al. (2017),we demonstrate that a state-of-the-art technique for domain adaptation, Optimal Transport, can be used to improve aphasia detection in a cross-language evaluation setting.

2. Background & Related Work

Existing studies have considered machine learning (ML) based approaches to aphasia detection Fraser et al. (2014); Le et al. (2017), but a summary of previous work reveals these have been largely restricted to single language settings (Tab. 1). While Fyndanis and Themistocleous (2018) considers multiple languages, only non-speech-related features were used and there was no model transfer across languages. To the best of our knowledge, there is no prior work where cross-language transfer has been used in the detection of aphasia.

Table 1: Summary of aphasia analysis in various languages, including usage of any external corpus in addition to the N reported (indicative of training/test size), feature modalities and performance

Methods	Language(s)	Features	Subjects/Samples (N,M)	External corpus (Y/N)	Performance
Fraser et al. (2014)	English	Linguistic	26, 26	N	1.00 (Accuracy)
Fyndanis and Themistocleous (2018)	German, Italian and Greek	Linguistic, cognitive performance indices	26, 104	N	0.79 (AUC)
Law et al. (2018)	Cantonese	Lexical and semantic	65, 65	N	-
Qin et al. (2018)	Cantonese	Text and acoustic	82, 328	N	0.90 (F1-score)
Ishkhanyan et al. (2017)	French	Lexical	15, 45	N	-
Ours	French	Text	24, 24	Y	0.87 (F1-score)
Ours	Mandarin	Text	60, 55	Y	0.69 (F1-score)

2.1. Domain Adaptation

Optimal Transport for Domain Adaptation: Diverse methods have been explored for domain adaption. Methods involving adversarial loss have been developed for multi-lingual word embeddings and language translation Conneau et al. (2018); Xu et al. (2018); some of them specialized to clinical machine translation Weng and Szolovits (2018). We use Optimal Transport, an embedding-based method of domain adaptation (Sec. 4.1). Variants of the base OT approaches have been proposed for a variety of NLP tasks in prior work Chen et al. (2019); Courty et al. (2017), including aligning representations across domains

in an unsupervised manner Bhushan Damodaran et al. (2018). OT-based sequence-to-sequence learning techniques have outperformed strong baselines in machine translation and abstractive summarization Chen et al. (2019), while modifications to the base OT algorithms have set new benchmarks for unsupervised word translation Alvarez-Melis et al. (2019).

Cross-linguistic Adaptation: Recent work in dementia detection, rather than aphasia, used paired samples from the OpenSubtitles Lison and Tiedemann (2016) dataset to train a regression model between independently engineered features from Mandarin and English transcripts Li et al. (2019). In contrast, we use unpaired data and learn mappings between *distributions* of the same linguistic features between different source languages and English. We utilize unpaired datasets in our study since this approach is more general and more useful when paired datasets are not available between a resource-rich and other languages. OT overcomes the requirement of paired data by aligning probability distribution functions of the linguistic features, rather than the features themselves.

3. Data Sources and Pre-processing

In this section, we provide details regarding all our data sources and text preprocessing steps.

3.1. AphasiaBank

All datasets of speakers of English, French and Mandarin are obtained from Aphasia-Bank [1] MacWhinney et al. (2011). The aphasic speakers have various subtypes of aphasia - broca, wernicke, anomic, etc. (See App. A) All participants perform multiple speech-based tasks, such as describing pictures, story-telling, free speech and discourse. We combine all tasks to a single transcript in our analysis. Detailed statistics for each language are in Table 2. All samples are manually transcribed, following the CHAT protocol Ratner (1993). We classify speech samples to two classes - healthy and aphasic, where aphasic constitutes all sub-types mentioned above, using extracted linguistic features.

3.2. TED Talks

We use a large dataset of TED talks with multi-lingual transcripts Tiedemann (2012) to train our domain adaption systems. In total, there are recordings available for 1178 talks, with various speaker accents and styles. We use transcripts from Mandarin, French and English languages. So that our domain adaption system is not biased by seeing paired data, which is not present in our aphasia classification task, we ensure there is no overlap between speech transcripts of English and French/Mandarin by dividing the talks into two sets and ensuring that the English transcripts for training the domain adaptation models are obtained from the first set, while those for French/Mandarin are obtained from the second set.[2] Additionally, similar to the methodology of Li et al. (2019), we attempt to create a larger dataset by dividing each narration into segments by considering 25 consecutive

---

1. https://aphasia.talkbank.org/

2. We also performed experiments to validate that this choice does not significantly affect results, finding that using either fully paired or fully unpaired (as described here) data yields statistically insignificantly different results.

utterances as one segment. We choose 25 because we observed that the features stabilize with this number of utterances (see App. B for details).

Table 2: Number of samples from AphasiaBank and the TED Talks corpus. Number of participants indicated in parentheses.

Corpus	Language	Healthy samples	Aphasic samples
AphasiaBank	English	246 (192)	428 (301)
AphasiaBank	French	13 (13)	11 (11)
AphasiaBank	Mandarin	42 (40)	18 (15)
TED Talks	English	2875 (589)	-
TED Talks	French	2976 (589)	-
TED Talks	Mandarin	2742 (589)	-

3.3. Transcript Pre-processing and Feature Extraction

The transcripts provided in AphasiaBank consist of transcribed speech following the CHAT protocol MacWhinney et al. (2011). Hence, there are several annotations such as repetitions, markers for incorrect word usage etc. To extract features, an important pre-processing step is to remove these various additional annotations. We utilize the *pylangacq* Lee et al. (2016) library for this step, due to its capabilities of handling CHAT transcripts. Additional pre-processing steps include stripping the various utterances of punctuations before POS-tagging. We extract the proportion of 8 POS[3] – nouns, verbs, subordinating conjunctions, adjectives, adverbs, coordinating conjunctions, determiners and pronouns – over the whole transcript of speech. Though aphasic speakers perform one additional speech task (where they provide details regarding their stroke) more than control speakers, these 8 features are agnostic to total length and content of transcripts, and rely more on the sentence complexity. These simple features are used because they are general and have been identified to be important in prior work Fraser et al. (2013); Li et al. (2019); Law et al. (2013) across languages. These features are extracted from all languages in AphasiaBank.

To analyse the variance in features across languages, we study if they differ significantly between healthy and aphasic speakers across languages (Tab. 5 in the App.). We observe that every feature varies significantly between healthy and aphasic speakers of English and Mandarin. We anticipate, hence, that raw, non-adapted cross-language transfer of models trained on English speech to Mandarin would lead to low performance.

4. Methods

We describe the domain adaptation system, which uses Optimal Transport. Overall pipeline in Fig. 1.

4.1. Cross-linguistic Representation Learning with Optimal Transport

Optimal transport (OT) consists of finding the best transport strategy from one probability distribution function (PDF) to another. This is done by minimizing the total cost of transporting a sample from the source to that in the target. Thus, there needs to be a metric
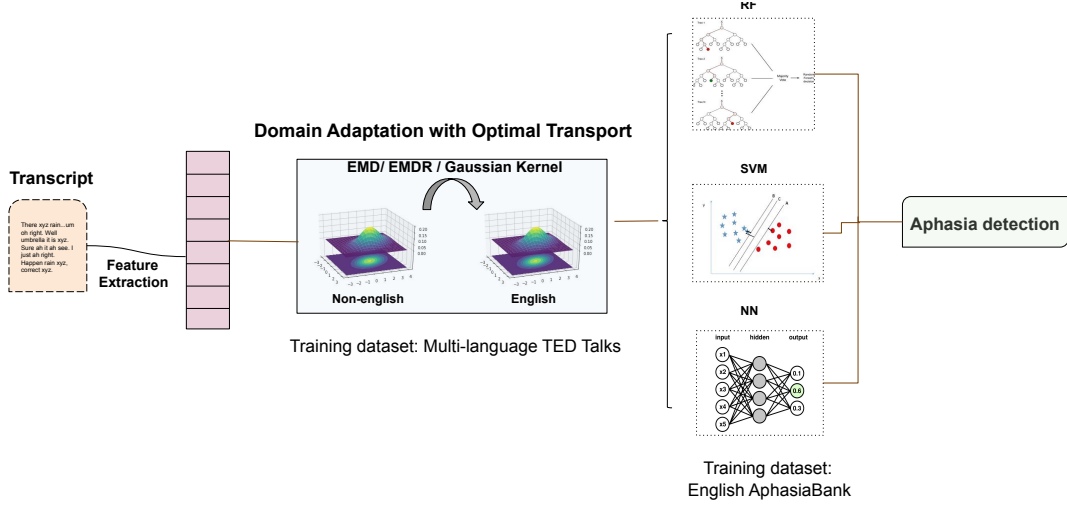
---

3. https://universaldependencies.org/u/pos/

Figure 1: Pipeline for processing a speech transcript from a non-English language. Features are extracted as detailed in Sec. 3.3, cross-lingual representations are obtained with multiple Optimal Transport algorithms, and aphasia detection classification performed with different ML models.

to quantify the different distances between samples in the two probability distributions, as well as solvers to solve the optimization problem of minimizing the total cost of transport, where cost is related to distance between source and target. We use optimal transport for domain adaptation here because we extract the same features across languages, though their distributions (in terms of feature values, e.g. proportion of nouns) vary from one language to another.

We use three solvers and distance functions between PDFs based on optimal transport:

Earth Movers Distance OT (EMD) EMD or Wasserstein distance between the two distributions is minimized using an optimal transport Network Flow Bonneel et al. (2011).

Gaussian Optimal Transport Mapping (Gaussian kernel) The Earth Movers Distance (EMD) or Wasserstein distance between the two distributions is minimized same as in 4.1. However, the transport map is approximated with a gaussian kernelized mapping to obtain smoother transport maps Perrot et al. (2016).

Entropic Regularization OT solver (EMD-R) Optimal transportation problem with EMD regularized by an entropic term, turning the linear program into a strictly convex problem that can be solved with the Sinkhorn-Knopp matrix scaling algorithm Sinkhorn and Knopp (1967). Linear solver proposed by Cuturi (2013) is used.

We employ open-source implementations of these algorithms Flamary and Courty (2017). We will refer to each of these OT algorithms as *EMD, Gaussian kernel, EMD-R* respectively as above. Detailed hyperparameter settings for early stopping tolerance, regularization terms (in EMD-R) are are in Appendix D.

OT mappings from each source language (French/Mandarin) to English are learned for each algorithm, trained on language pairs (English-French/English-Mandarin) from the TED Talks dataset.

5. Experiments

We consider the classification of speech as aphasic or healthy from speech transcripts (as featurized via POS proportions), and we are primarily interested in whether performance can be improved for low-resource languages. This classification task is performed across several baseline settings, including a unilingual task, direct feature transfer, and an autoencoder-mediated multilingual encoding, as well as various domain adaptation settings (Sec. 5.2), where features are mapped from low resource languages to English using Optimal Transport. A suite of ML models, including SVM, RF and NN are used for this task (see Sec. 5.4 for hyperparameters).

We evaluate task performance primarily using macro-averaged F1 scores, a measure that is known to be robust to class imbalance. We also report AUROC scores, since it is often used as a general measure of performance irrespective of any particular threshold or operating point Richardson and Domingos (2006); Liu and Shriberg (2007).

Due to the lack of baselines on the multilingual AphasiaBank dataset in prior work, we establish our own unlingual and multilingual baselines, detailed in the section below.

5.1. Baseline Domain Adaptation Systems

Unilingual Training: Unilingual baselines for each language using 10-fold cross-validation (CV), stratified by subject so that each subject's samples do not occur in both training and testing sets in each fold. We estimate that this would be a lower bound on performance for French and Mandarin AphasiaBank, since it is likely, given the small size of the dataset, that models would underfit and have low generalizable performance across subjects.

Feature Transfer from English with raw, non-adapted features: We also identify transfer baselines, wherein models trained on English AphasiaBank are evaluated on other languages with no fine-tuning. We hypothesize that this baseline would be more performant than the uni-lingual baseline, at least amongst the more-similar Romance languages of French and English, since it utilizes the comparatively larger dataset of English AphasiaBank for training.

Multilanguage Embedding with an Autoencoder: A common representation is obtained for all three languages by encoding the linguistic features using a high capacity autoencoder. This autoencoder, trained on English, French and Mandarin TED Talks datasets (unpaired), maps linguistic features extracted from multilingual transcripts into a shared latent space. The autoencoder consists of 4 hidden layers (2 hidden layers in encoder and decoder respectively) with 5, 3, 3 and 5 units each for the following experiment. Hyperparameters are set using a 90-10 train-dev split of samples from each language. All ML classifiers then are trained on the encoded versions of English AphasiaBank, and tested on encoded versions of French and Mandarin AphasiaBank. Comparison of other training regimes to this baseline would determine if learning a shared representation across multiple languages is better than OT.

5.2. Training Regimes for OT Adapted-transfer

We evaluate two OT-training regimes for both French and Mandarin, each tested across all varieites of OT (e.g., OT-EMD, OT-Gaussian, and OT-EMD-R).

Feature Transfer from English with OT domain adaptation, with TED Talks for OT: Models trained on English AphasiaBank evaluated on other languages with OT adaptation (EMD, EMD-R, Gaussian mapping) with no fine-tuning. OT models here are trained only on the multi-language unpaired TED Talks dataset, i.e, *with no aphasic data*.

Feature Transfer from English with OT domain adaptation, with TED Talks and AphasiaBank for OT: Models trained on English AphasiaBank evaluated on other languages with OT adaptation (with EMD, EMD-R, Gaussian mapping) with no fine-tuning. The OT models here are trained on the multi-language TED Talks dataset, and multi-language AphasiaBank i.e, *with aphasic data*. We ensure that there is no overlap in the proportion of AphasiaBank used for learning OT mappings and for evaluating the classifiers by employing 2-fold cross-validation where one fold is included in the training set for OT and another for evaluation. Since OT involves source and target domain probability estimation, we hypothesize that adding in-domain data, particularly that of speech-impaired participants, would improve results significantly.

5.3. Impact of Including Diverse Speech Samples in OT Training

Since literature shows that accents can have a significant effect on POS features Runnqvist et al. (2013), we hypothesize that there is an observable effect of diversity in terms of accents in the OT training set. To study this effect, we manually annotate accents for the English TED Talks dataset (details in App. G) as 'North American'(NA) accent or 'other' accent. In total in the TED Talks English set, there are 373 NA accented and 215 'other' accented talks. We study the impact of increasing the diversity of accents used for training OT algorithms, keeping the size of the dataset constant (see Tab. 4).

5.4. Hyperparameter Settings

Hyperparameters for classification models are tuned using grid search with 10-fold cross validation on the training set (English AphasiaBank) across all settings. We use an SVM (RBF kernel with regularization parameter $C = 0.1$ and $\gamma = 0.001$), Random Forest (RF; 200 decision trees with a maximum depth of 2), and Neural Network (NN; 2 hidden layers of 100 units each) classifiers for the cross-linguistic classification taskPedregosa et al. (2011). Since the training set is highly imbalanced (see Tab. 2), the minority class is oversampled synthetically with SMOTE Chawla et al. (2002) with $k = 3$. Prior to oversampling, the training set is normalized and scaled using the median and interquartile range, a common mechanism to center and scale-normalize data which is robust to outliers Pedregosa et al. (2011). The same median and interquartile (obtained from the training set) bounds are used to scale the evaluation set in each case.

All ML classifiers are trained completely only on English AphasiaBank, while the OT models are trained on unpaired samples across English and another language (French or Mandarin) from the TED talks corpus Tiedemann (2012).

6. Results

In Tab. 3, we compare the performance of various OT-algorithms to their respective baselines for cross-language representation learning for the aphasia detection prediction task.

Baselines We see that baseline performance varies significantly between languages. For French, using a multilingual encoding or direct feature transfer largely offers significant improvements over unilingual training, yielding a maximal lift of 15 for RF mean F1, and achieving maximal overall classifier performance using multilingual encoding with an SVM model. In general for French, the multilingual encoding outperforms the feature transfer baseline, but both improve on unilingual results.

For Mandarin, these results are very different; here, either baseline approach to adaption hurts overall performance as compared to a unilingual baseline, often yielding solutions for which the model will predict just a single class exclusively.

Table 3: F1 macro and AUROC mean and standard deviation scores across languages for different model settings in OT training, averaged across multiple runs. Note that zero standard deviations occur when a single class is predicted or when standard deviation < 0.01. The standard deviations are an artefact of the small sample sizes in the evaluation set (24 and 60 for French and Mandarin respectively), as seen in prior literature Fraser et al. (2013). Highest F1 scores are shown in bold for each language and classifier. Overall, the highest mean F1 scores are obtained with OT adaptation with the EMD variants, with aphasic samples included in OT training, along with the multilingual TED Talks dataset. . A study on the effect of data size on uni-lingual performance in English is in App. F for comparison.

Language	Method	SVM		RF		NN	
		F1	AUROC	F1	AUROC	F1	AUROC
French	Unilingual Baseline	74.00 ± 0.00	80.00 ± 0.00	64.00 ± 5.44	72.50 ± 4.08	76.67 ± 0.00	79.17 ± 2.30
	Mutlilingual Encoding	$\mathbf{85.58 \pm 3.79}$	$\mathbf{85.52 \pm 4.07}$	79.57 ± 5.25	79.44 ± 4.90	81.53 ± 4.70	81.96 ± 4.72
	Feature Transfer	79.13 ± 0.00	78.61 ± 0.00	77.23 ± 0.00	77.27 ± 0.00	52.93 ± 5.04	58.97 ± 1.81
	OT-EMD	79.13 ± 0.00	79.38 ± 0.00	80.49 ± 1.92	81.12 ± 2.57	65.18 ± 1.94	64.80 ± 2.14
	OT-Gaussian	53.13 ± 0.00	61.54 ± 0.00	41.89 ± 3.38	56.41 ± 1.81	39.50 ± 0.00	53.85 ± 0.00
	OT-EMD-R	83.22 ± 0.00	83.22 ± 0.00	81.76 ± 2.07	81.70 ± 2.14	73.48 ± 1.91	74.94 ± 1.81
	OT-EMD - with aphasic	83.10 ± 0.00	84.52 ± 0.00	80.26 ± 2.46	82.14 ± 2.06	78.84 ± 0.00	80.95 ± 0.00
	OT-Gaussian - with aphasic	45.96 ± 0.00	58.33 ± 0.00	39.50 ± 0.00	54.17 ± 0.00	39.50 ± 0.00	54.17 ± 0.00
	OT-EMD-R - with aphasic	$\mathbf{87.23 \pm 0.00}$	$\mathbf{88.09 \pm 0.00}$	$\mathbf{83.10 \pm 0.00}$	$\mathbf{84.52 \pm 0.00}$	$\mathbf{81.68 \pm 2.45}$	$\mathbf{83.33 \pm 2.07}$
Mandarin	Unilingual Baseline	60.47 ± 0.00	57.92 ± 0.00	57.78 ± 0.54	60.08 ± 1.02	57.66 ± 2.91	57.19 ± 1.22
	Mutlilingual Encoding	23.08 ± 0.00	50.00 ± 0.00	30.92 ± 13.38	51.19 ± 4.12	23.08 ± 0.00	50.00 ± 0.00
	Feature Transfer	23.08 ± 0.00	50.00 ± 0.00	23.08 ± 0.00	50.00 ± 0.00	23.08 ± 0.00	50.00 ± 0.00
	OT-EMD	63.28 ± 0.00	67.06 ± 0.00	55.41 ± 0.06	57.01 ± 0.67	53.79 ± 1.32	59.92 ± 2.34
	OT-Gaussian	31.80 ± 0.00	51.98 ± 0.00	30.26 ± 1.09	51.19 ± 0.56	27.11 ± 0.00	49.60 ± 0.00
	OT-EMD-R	$\mathbf{66.25 \pm 0.00}$	67.46 ± 0.00	54.43 ± 1.85	58.33 ± 3.29	56.44 ± 2.20	61.11 ± 0.85
	OT-EMD - with aphasic	65.59 ± 0.00	$\mathbf{70.57 \pm 0.00}$	$\mathbf{69.05 \pm 1.34}$	$\mathbf{68.10 \pm 0.91}$	55.92 ± 2.84	59.00 ± 3.18
	OT-Gaussian - with aphasic	34.75 ± 0.00	54.32 ± 0.00	32.92 ± 0.00	53.18 ± 0.00	26.82 ± 0.00	49.77 ± 0.00
	OT-EMD-R - with aphasic	59.32 ± 0.00	59.09 ± 0.00	61.18 ± 0.83	60.23 ± 0.99	$\mathbf{62.57 \pm 0.12}$	$\mathbf{64.39 \pm 0.06}$
English	Unilingual Baseline	85.89 ± 0.00	88.93 ± 0.00	82.14 ± 0.31	85.01 ± 0.11	88.07 ± 0.35	88.49 ± 0.52

OT-Variants Among OT-Variants, we see generally stronger performance as compared to the unilingual models and baseline domain adaption systems as well. In all but one case, the best-performing OT- variant for a given model/language yields a statistically significant improvement over the best baseline model according to a paired *t*-test, the notable exception being for the SVM model on French text, which does not achieve statistical significance. In general, EMD variants of OT (including both OT-EMD-R and OT-EMD) tend to perform

better than OT-Gaussian, and nearly universally, including aphasic speech samples in the OT model yields significant lifts, yielding best in class mean-F1 of 87.23 for an SVM model over French samples under the OT-EMD-R model, or 69.05 for a RF model over Mandarin text via the OT-EMD model.

Speech Diversity   We additionally analyze how speech diversity, as measured by frequency of various accents in the speech data, affects the performance of OT-EMD-R domain adaption for SVM models in French and Mandarin. Results for this are shown in Table 4. For both French and Mandarin in this case, we observe that increasing the prevalence of non-North American accents in the domain adaption task improves downstream aphasia/non-aphasia classification performance by several F1 points (yielding a score of 87.48 for French and 69.19 for Mandarin). Note that these results are not statistically significantly different than the best results found previously in Tab. 3.

Table 4: F1 macro scores across languages with OT-EMDR, with varying proportions of data. Highest scores are shown in bold. Note that we don't report standard deviation since it < 0.01 in all cases.

Language	Method	OT Dataset Size	SVM	
			F1	AUROC
French	OT-EMD-R	286 NA	83.22	83.22
	OT-EMD-R	215 NA, 71 not NA	83.22	83.22
	OT-EMD-R	143 NA, 143 not NA	83.22	83.22
	OT-EMD-R	71 NA, 215 not NA	**87.48**	**87.76**
Mandarin	OT-EMD-R	286 NA	66.25	67.46
	OT-EMD-R	215 NA, 71 not NA	62.50	63.49
	OT-EMD-R	143 NA, 143 not NA	68.51	70.24
	OT-EMD-R	71 NA, 215 not NA	**69.19**	**71.83**

7. Discussion

Direct Feature Transfer Only Relevant in Similar Languages   In Tab. 3, we observe that direct feature transfer (i.e., the "Feature Transfer" row) achieves good performance for the English to French domain adaption task, but not for the English to Mandarin adaption task. This makes sense as English and French have relatively similar grammatical patterns Roberts (2012) (e.g., subject, verb, object ordering) whereas Mandarin and English have a number of significant differences, including, e.g., reduplication, where a syllable or word is repeated to produce a modified meaning, in Mandarin Li and Thompson (1989).

Relatedly, the multilingual encoding approach likewise yields good performance only for French. Here, we again note that French and English are relatively similar languages, compared to English and Mandarin. Thus, our multilingual encoder may be much more able to jointly encode English and French than it could English and Mandarin.

Inclusion of Aphasic Samples is Highly Impactful on OT-performance   We observe (Tab. 3), that the highest mean F1-score for cross-language classification on the evaluation set increases to 87.23 (OT-EMD-R with SVM) for French and 69.04 (OT-EMD with SVM) for Mandarin from 83.22 and 66.25 respectively (both significant increases, with $p < 0.001$ and $p = 0.015$ respectively) with the addition of aphasic samples in the training set for OT adaptation. This demonstrates that including aphasic samples has a strong

positive effect on OT- based domain adaption. Note that similar results of performance improvement due to addition of in-domain, speech-impaired data have also been observed for multi-lingual topic modelling from speech in prior literature Fraser et al. (2019).

Diverse Speech Samples in Representation Improve Performance   As stated in Section 6, we find that increasing the diversity of our OT dataset (as measured through accent distribution) has a positive effect on downstream transfer. This resonates with prior findings, which have shown that accents can have a significant effect on POS features Runnqvist et al. (2013).

8. Conclusions and Future Work

A limitation of our current work is that it focuses mainly on a single method of domain adaptation, Optimal Transport. Additionally, the feature set is limited to only include text-based features and hence, results are dependant on the features extracted, and change in feature space might have a significant effect on the relative performance of domain adaptation and multilingual representation learning. In future work, we will empirically compare OT domain adaptation strategy with other techniques, such as adversarial domain adaptation Tzeng et al. (2017) for the aphasia detection task. Furthermore, we plan to study the effect of different featurizations (such as voice acoustics and inclusion of more linguistic features.) on overall performance with the current setup.

Availability of datasets of an appropriate quality and size is essential in the ML for Healthcare domain, due to the high cost of errors, as well as to ensure fair decisions for all individuals Rajkomar et al. (2018). Various solutions have been proposed previously for mitigating the problem of data availability, including creating novel sources of data, developing data-efficient algorithms, and employing domain adaptation from low-resource to resource-rich domains Li et al. (2019); Bull et al. (2018). However, the importance of standard, diverse, in-domain medical speech datasets is underscored by our observations made in Sec. 7.

In summary, we show that POS features extracted from speech transcripts from different languages can be mapped to English to aid in clinical speech classification task. We find that the OT strategy is successful in domain adaptation, with associated increase in classification performance for French and Mandarin over unilingual baselines. In comparison to a multilingual baseline with a high-capacity autoencoder, OT algorithms work on par for similar, and significantly better for dissimilar languages. Our results suggest that domain adaption strategies, in particular OT-based domain adaption, can help enable strong predictive models for aphasia detection in low-resource languages

Acknowledgments

References

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879, 2019.

Elizabeth Bates, Beverly Wulfeck, and Brian MacWhinney. Cross-linguistic research in aphasia: An overview. *Brain and language*, 41(2):123–148, 1991.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *ACM Transactions on Graphics (TOG)*, volume 30, page 158. ACM, 2011.

L Bull, K Worden, G Manson, and N Dervilis. Active learning for semi-supervised structural health monitoring. *Journal of Sound and Vibration*, 437:373–388, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. In *International Conference on Learning Representations*, 2019.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017. URL https://github.com/rflamary/POT.

Kathleen C Fraser, Frank Rudzicz, and Elizabeth Rochon. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181, 2013.

Kathleen C Fraser, Jed A Meltzer, Naida L Graham, Carol Leonard, Graeme Hirst, Sandra E Black, and Elizabeth Rochon. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, 55:43–60, 2014.

Kathleen C Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139, 2019.

Valantis Fyndanis and Charalambos Themistocleous. Morphosyntactic production in agrammatic aphasia: A cross-linguistic machine learning approach. *Frontiers in Human Neuroscience*, (75), 2018. ISSN 1662-5161. doi: 10.3389/conf.fnhum.2018. 228.00075. URL http://www.frontiersin.org/human_neuroscience/10.3389/conf. fnhum.2018.228.00075/full.

Byurakn Ishkhanyan, Halima Sahraoui, Peter Harder, Jesper Mogensen, and Kasper Boye. Grammatical and lexical pronoun dissociation in french speakers with agrammatic aphasia: a usage-based account and ref-based hypothesis. *Journal of Neurolinguistics*, 44:1–16, 2017.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.

Susanne Kristen, Sabrina Chiarella, Beate Sodian, Tiziana Aureli, Maria Genco, and Diane Poulin-Dubois. Crosslinguistic developmental consistency in the composition of toddlers' internal state vocabulary: Evidence from four languages. *Child Development Research*, 2014, 2014.

S Law, A Kong, L Lai, and C Lai. Production of nouns and verbs in picture naming and narrative tasks by chinese speakers with aphasia. *Procedia, social and behavioral sciences*, 94, 2013.

Sam-Po Law, Anthony Pak-Hin Kong, and Christy Lai. An analysis of topics and vocabulary in chinese oral narratives by normal speakers and speakers with fluent aphasia. *Clinical linguistics & phonetics*, 32(1):88–99, 2018.

Duc Le, Keli Licata, and Emily Mower Provost. Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298, 2017.

Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. Working with chat transcripts in python. Technical Report TR-2016-02, Department of Computer Science, University of Chicago, 2016.

Bai Li, Yi-Te Hsu, and Frank Rudzicz. Detecting dementia in mandarin chinese using transfer learning from a parallel corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1991–1997, 2019.

Charles N Li and Sandra A Thompson. *Mandarin Chinese: A functional reference grammar*. Univ of California Press, 1989.

Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. 2016.

Yang Liu and Elizabeth Shriberg. Comparing evaluation metrics for sentence boundary detection. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–185. IEEE, 2007.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011.

NAA. Aphasia fact sheet, 2016. URL https://www.aphasia.org/aphasia-resources/aphasia-factsheet/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205, 2016.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL https://nlp.stanford.edu/pubs/qi2018universal.pdf.

Ying Qin, Tan Lee, and Anthony Pak Hin Kong. Automatic speech assessment for aphasic patients based on syllable-level embedding and supra-segmental duration features. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5994–5998. IEEE, 2018.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 2018.

Nan Bernstein Ratner. Brian macwhinney, the childes project: Tools for analyzing talk. hillsdale, nj: Erlbaum, 1991. pp. xi+ 360. *Language in Society*, 22(2):307–313, 1993.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62 (1-2):107–136, 2006.

Ian G Roberts. *Verbs and diachronic syntax: A comparative history of English and French*, volume 28. Springer Science & Business Media, 2012.

Elin Runnqvist, Tamar H Gollan, Albert Costa, and Victor S Ferreira. A disadvantage in bilingual sentence production modulated by syntactic frequency and similarity across languages. *Cognition*, 2(129):256–263, 2013.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Efstathia Soroli, Halima Sahraoui, and Carol Sacchett. Linguistic encoding of motion events in english and french. 2012.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

Wei-Hung Weng and Peter Szolovits. Mapping unparalleled clinical professional and consumer languages with embedding alignment. *KDD Workshop on Machine Learning for Medicine and Healthcare*, 2018.

Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. Unsupervised clinical language translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 3121–3131, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330710. URL http://doi.acm.org/10.1145/3292500.3330710.

Jingjing Xu, Xu SUN, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia, jul 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P18-1090.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298, 2018.

Appendix A. Aphasia sub-types

- Broca aphasia or non-fluent aphasia: Individuals with Broca's aphasia have trouble speaking fluently but their comprehension can be relatively preserved.

- Wernicke's aphasia or fluent aphasia: In this form of aphasia the ability to grasp the meaning of spoken words is chiefly impaired, while the ease of producing connected speech is not much affected.

- Anomic aphasia: Individuals with anomic aphasia can understand speech and read well but frequently are unable to obtain words specific to what they wish to talk about – particularly nouns and verbs.

- Transcortical aphasia: Individuals with this type of aphasia have reduced speech output, typically due to a stroke.

- Conduction aphasia: Individuals with can comprehend speech and read well, but have significant difficulty in repeating phrases.

Appendix B. Choosing Transcript Length from TED Talks

We compare differences in values of the 8 POS speech features from speech samples, for transcript lengths of 5, 25, 50, 75 and 100 utterances each. We compute t-tests between features computed from transcripts lengths of 5 and 25, 25 and 50, 50 and 75, and 75 and 100. We find that while features 5 out of 8 features are significantly different between transcript lengths of 5 and 25, they stabilize for lengths greater than or equal to 25, i.e, no significant difference between lengths of 25 and 50 (lowest p-value is 0.22), 50 and 75 (lowest p-value is 0.32) and 75 and 100 (lowest p-value is 0.59). Thus, we choose 25 utterances to be the standard length of a transcript from the TED Talks dataset to maximize data available.

Appendix C. Part-of-Speech Proportions Comparisons Across Languages

Table 5: Significant p-values corresponding to T-tests of the 8 features between English and other languages (after Bonferroni correction). Indicated by '*' if significantly different for both Mandarin and French, '+' if only significantly different between English and Mandarin, '#' if only significantly different between English and French and '-' if there is no significant difference.

POS/Feature	Aphasia	Control
Nouns	+	+
Verbs	*	*
Subordinating conjunctions	*	*
Adjectives	+	*
Adverbs	*	*
Co-ordinating Conjunctions	+	*
Determiners	+	*
Pronouns	+	+

Appendix D. Hyperparameters

For EMD, method proposed by Ferradans et al. (2014) is used for out of sample mapping to apply to transport samples from a domain into the other with other default parameters in Flamary and Courty (2017).

For EMD-R, entropic regularization parameter is set to 3 with all other parameters default.

For Gaussian mapping, the weight for linear OT loss is set to 1, and maximum iterations is set to 20, with stop threshold for iterations set to $1e - 05$ with other default parameters in Flamary and Courty (2017).

Appendix E. Paired Data Does Not Improve Performance Significantly

We observe, from Tab. 6, that paired data does not significantly improve performance for French or Mandarin, over classification with unpaired datasets for OT.

Table 6: F1 macro scores across languages with OT, with paired data.

Language	Method	SVM	RF	NN
		F1	F1	F1
French	OT-EMD	$\mathbf{83.22 \pm 0.00}$	$\mathbf{84.71 \pm 1.95}$	67.22 ± 2.65
	OT-Gaussian	46.67 ± 0.00	41.89 ± 3.38	34.12 ± 3.80
	OT-EMD-R	$\mathbf{83.22 \pm 0.00}$	78.84 ± 0.00	74.97 ± 3.41
Mandarin	OT-EMD	59.28 ± 0.00	$\mathbf{55.35 \pm 3.38}$	49.30 ± 2.42
	OT-Gaussian	25.97 ± 0.00	25.53 ± 0.63	24.64 ± 0.00
	OT-EMD-R	$\mathbf{60.11 \pm 0.00}$	49.25 ± 1.26	$\mathbf{56.12 \pm 2.01}$

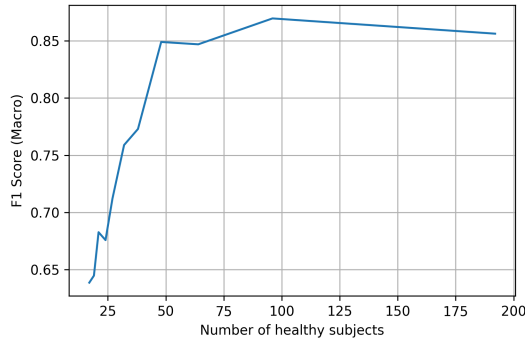## Appendix F. Studying Effect of Data on Unilingual Performance

Figure 2: Effect of dataset size on the aphasia detection task.

To study the impact of data on the aphasia detection task, we perform an ablation study wherein the size of the English AphasiaBank dataset is artificially reduced by integer factors (while keeping the relative proportion of healthy and aphasic subjects same). We

perform 10-fold cross-validation for a SVM classifier, with progressively less data. We observe that speech transcripts from atleast 50 healthy subjects are required for the classification performance to stabilize, given the current feature set (see Fig. 2). F1 scores (micro and macro) increase non-linearly with the addition of data.

Appendix G. Accent Annotation

The TED-Talks dataset covers a wide speaker demographic, in terms of sex, age and accents. Since prior literature shows that accents can have a significant effect on linguistic features Runnqvist et al. (2013), we manually annotate presence or absence of a North-American accent for English speech in the TED-Talks dataset. An annotator listens to the audio associated with each TED-Talk and annotates if the accent is 'North American' or not. In cases where the accent is not clear, publicly available information regarding nationality of speaker is referenced. In future work, we plan to have multiple annotations per audio, and factor in metrics such as cross-rater agreement into our analysis.