Towards Domain-Agnostic and Domain-Adaptive Dementia Detection from Spoken Language

Shahla Farzana and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois Chicago
{sfarza3, parde}@uic.edu

Abstract

Health-related speech datasets are often small and varied in focus. This makes it difficult to leverage them to effectively support healthcare goals. Robust transfer of linguistic features across different datasets orbiting the same goal carries potential to address this concern. To test this hypothesis, we experiment with domain adaptation (DA) techniques on heterogeneous spoken language data to evaluate generalizability across diverse datasets for a common task: dementia detection. We find that adapted models exhibit better performance across conversational and task-oriented datasets. The feature-augmented DA method achieves a 22% increase in accuracy adapting from a conversational to task-specific dataset compared to a jointly trained baseline. This suggests promising capacity of these techniques to allow for productive use of disparate data for a complex spoken language healthcare task.

1 Introduction

Data-driven models of diverse conditions affecting spoken language abilities offer promising real-world benefits (Amini et al., 2022; Girard et al., 2022). However, the datasets available for developing these models are often small and disparate, spanning varied diagnostic and non-diagnostic tasks mapped to different taxonomies at conflicting granularities (Graham et al., 2020). This has largely constrained progress to models excelling in specialized settings (e.g., individuals with homogeneous language background describing a standardized image (Luz et al., 2020)). At the same time, it has created challenges in building more generalizable knowledge about language patterns associated with the condition of interest (Guo et al., 2021).

Outside healthcare applications, *domain adaptation* (DA) has long been applied to increase the capacity of NLP systems to leverage meaningful information from diverse data (Kouw and Loog, 2018). These techniques generally seek to harness data from one domain (the *source*) to improve performance in another (the *target*). Usually the target domain has little or no labeled data, while the source has a relatively large amount of labeled data. Despite the advantages offered by DA for many NLP problems, it has remained under-studied for healthcare tasks due to numerous complexities of healthcare data (Laparra et al., 2020). Nonetheless, most healthcare problems offer the ideal learning settings in which DA is designed to thrive.

We present a systematic analysis of the use of DA for a low-resource healthcare problem that has recently been popular in the NLP community: *dementia*. We adopt a wide definition of dementia in our work, encompassing datasets pertaining to Alzheimer's disease or related dementia (ADRD) and age-related mild cognitive impairment (MCI), in line with current NLP community norms (Amini et al., 2022). Our research questions include:

Q1. Can DA be used to exploit spoken language data pertaining to dementia from one domain, to improve its detection in other domains?

Q2. If yes, does this offer performance improvements over simpler joint training?

Q3. How do different linguistic features and class biases contribute to this performance?

We define *domain* in this study as a distinct dataset with supervised labels describing dementia status in some capacity. Data collection protocol and precise labeling taxonomy may vary across domains, making our task slightly more complex than related work that focused solely on differences in source language (Balagopalan et al., 2020b) or labeling taxonomy (Masrani et al., 2017). We find that DA can indeed support improved dementia detection across domains compared to joint training, and we identify key pivot features and factors

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11965–11978

contributing to this success. It is our hope that continued study of DA in healthcare applications can further extend the boundaries of our understanding and promote impactful follow-up work.

2 Related Work

Most prior work on developing spoken language models of dementia has followed a common pattern, focusing on designing and evaluating dataset-specific approaches. This has included (most popularly) a picture description task (Balagopalan et al., 2020a; Yuan et al., 2020; Di Palo and Parde, 2019), as well as other datasets with more open-ended conversational speech (Li et al., 2022; Nasreen et al., 2021b; Luz et al., 2018). These models singularly focus on the source domain, with no expectation of deployment beyond that, opening questions about their ability to generalize beyond small, publicly available reference sets.

The extent to which DA has been explored in this context is limited. Li et al. (2022) leveraged transfer learning, one form of DA that involves fine-tuning a model pretrained on a much larger dataset using the smaller target domain dataset, to study the perplexity ratio of normal and artificially degraded Transformer-based language models for dementia detection. Likewise, Balagopalan et al. (2020b) achieved performance boosts in detecting early signs of *aphasia* in cross-language settings compared to the unilingual baseline using optimal transport domain adaptation. A problem with transfer learning in many healthcare contexts is that target datasets are much smaller than for other NLP tasks for which the technique has demonstrated success. The benefits of transfer learning do not necessarily transfer (no pun intended) to ultra low-resource settings, where resulting models may be much less stable (Dodge et al., 2020).

Other forms of DA that may be more suited to dementia detection and other very low-resource healthcare problems are feature-based and instance-based DA. Both were originally leveraged for smaller datasets closer in scale to (although still larger than) those available for dementia detection (Daumé III, 2007; Sun et al., 2016), making it a promising and perhaps under-appreciated alternative to transfer learning. Feature-based DA focuses on modifying the feature space of the source and target datasets in some way that promotes the classifier's ability to generalize across them. Masrani et al. (2017) experimented with two feature-based DA techniques to adapt separate domain subsets split from the same source dataset, DementiaBank (Becker et al., 1994). Instance-based DA focuses on reweighting instances based on their importance to the target domain task (Jiang and Zhai, 2007; Xia et al., 2014). It has not yet been studied for dementia detection. We build upon Masrani et al. (2017)'s promising findings by studying the effects of numerous feature-based and instance-based DA techniques across different dementia datasets with conversational and task-related speech samples.

3 Methodology

3.1 Task Definition

For the scope of the work presented here we abstract *dementia detection* to the following scenario. Given a dataset with instances X and labels Y from some domain D, then our label space $\mathbf{y} = \{d, c\} \in Y$ is drawn from the binary distribution of classes (e.g., {*probable Alzheimer's*, *control*} or {*with dementia*, *without dementia*}) present in D. We assign the class with an association most proximal to a dementia diagnosis (e.g., *possible Alzheimer's* or *with dementia*) to the dementia (d) label, and the other class to the control (c) label. Our goal is to predict $y_i \in Y$ for an unseen instance x_i with feature representation $\mathbf{x_i}$, which may be modified from the original representation according to the applied DA approach.

3.2 Data

We use three publicly available datasets and one privately-held dataset, representing separate domains, to study DA in this context. The publicly available datasets, DementiaBank, ADReSS, and the Carolinas Conversation Collection, are the most widely used datasets for dementia detection research in the NLP community. They are also the only datasets for which public access is available.[1] Characteristics of these datasets are provided in Table 1. In Figure 1, we provide samples from two of these datasets, quoted directly from Chinaei et al. (2017) and Davis et al. (2017), to illustrate language differences between task-oriented and conversational domains. Our privately-held dataset is used only for conditions requiring multiple source domains, explained in detail in §3.3.

[1] Researchers are still required to obtain permission from the dataset creators prior to using each of these datasets, via established processes that range from email request (Becker et al., 1994) to full review and approval by local and external Institutional Review Boards (Pope and Davis, 2011).

Dataset		# P	# T	L	SD
$ADReSS_d$	tr	54	54	125.5	81.8
	te	24	24	95.0	47.0
$ADReSS_c$	tr	54	54	134.7	59.4
	te	24	24	120.0	72.0
DB_d		162	243	124.8	67.9
DB_c		99	303	133.9	67.4
CCC_d		46	97	1320.7	1059.1
CCC_c		36	192	776.9	469.7
$ADRC_d$		3	3	444.7	132.6
$ADRC_c$		82	82	786.4	338.3

Table 1: Descriptive dataset characteristics. The subscripts *d* and *c* refer to *dementia* and *control*, respectively. *Length* (L) is provided as average number of words per transcript. DB and CCC have differing *# Participants* (P) and *# Transcripts* (T) because some participants in those datasets had multiple recorded interviews. ADReSS is subdivided into standardized *(tr)ain* and *(te)st* partitions established by the dataset's creators.

DementiaBank

INV: just tell me whats happening in the picture . .
PAR: the pearl [: poor] [* p:w] &mo moms gettin(g) her wet [/ /] feet wet (be)cause she thinking of days gone by and then the water run . [+ gram] .
PAR: (.) and &uh that boy whether he knows or not hes gonna [: going to] crack his head on the back of that counter trying to get too many cookies out . .

Carolinas Conversation Collection

INV: was it just (overlap)
PAR: um, my doctor was telling me all kind of little thingies, been so long, I forgot now. But, um, my nerves was bad.
INV: Your nerves?
PAR: Um hmm. And, um, I had a little heart failure. Um hmm. And, um, --- (long pause) that all, what else he tell me that was wrong? He say, "You got to stop," he just didn't tell me then, (overlap)

Figure 1: Characteristic language samples from DB (Chinaei et al., 2017) and CCC (Davis et al., 2017).

DementiaBank (DB). DB (Becker et al., 1994) is a publicly available compendium of audiorecordings of neuropsychological tests administered to healthy participants and patients with diagnosed dementia. It is the most widely used dementia detection dataset in the NLP community, and each audiorecording is paired with a manual transcription formatted using the CHAT transcription protocol (Macwhinney, 2009). We refer readers to Becker et al. (1994) for a detailed description of the dataset collection procedures and its overall composition.

The neuropsychological tests include a picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1972), often referred to as the "Cookie Theft Picture Description Task." Participants are presented with a picture stimulus which depicts numerous events, central to which is a boy stealing a cookie from a jar. They are asked to describe everything they see occurring in the picture. The bulk of the dementia detection work conducted using DementiaBank has focused on the English-language interactions from this task. DB contains 169 subjects with *probable Alzheimer's disease* and 99 *control* subjects.

Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS). ADReSS (Luz et al., 2021) is a subset of DB created for a series of shared tasks on dementia detection. Control and dementia subjects are matched in terms of age and gender, resulting in a balanced set of 156 samples (78 *with dementia* and 78 *controls*) split into training and test. The goal in developing ADReSS was to eliminate possible biases that may arise due to label and demographic imbalance in the original DB, at the expense of resulting in an ultimately smaller dataset. Its existence presents an interesting opportunity for comparison of balanced and unbalanced versions of the same source data. Since these datasets are drawn from the same source, we do not adapt DB to ADReSS or vice versa.

Carolinas Conversation Collection (CCC). CCC (Pope and Davis, 2011) is not derived from a neuropsychological task; instead, it focuses on English conversational speech. The dataset, collected by researchers studying language and healthcare across numerous institutions, contains 646 recorded interviews of 48 elderly cognitively normal individuals with non-dementia related conditions, and 284 individuals with dementia. Interview topics vary considerably. Members of the cohort *without dementia* have one interview with a young clinical professional and one with a demographically similar community peer, whereas members of the cohort *with dementia* have anywhere from 1-10 interviews with researchers and student visitors. The target focus of the conversational interviews is on eliciting autobiographical narrative pertaining to health and wellness. Although much less commonly used in the NLP community, it has recently been included a study that focus on the intersection

between interaction patterns and dementia status (Nasreen et al., 2021a), study regarding dementia-related linguistic anomalies in human language (Li et al., 2022), and so on. We used a transcribed subset of this corpus.

Alzheimer's Disease Research Center (ADRC). ADRC is a new, privately held dataset containing audiorecordings and matched transcriptions for a population of 85 elderly participants. Audiorecordings were collected during a structured narrative storytelling task, in which participants were asked to describe a memorable event from their young adulthood. Diagnoses were provided by trained psychiatrists. Audiorecordings were transcribed in a semi-automated manner, with an initial pass completed using the `Vosk`[2] speech recognition toolkit and a follow-up pass during which trained undergraduates manually corrected errors in the transcripts. Although not yet publicly available, plans are in place to release this dataset following guidelines created in concert with our psychiatric collaborators in an approved protocol from the Institutional Review Board at the University of California San Diego. We encourage interested parties to contact us for additional details.

3.3 Domain Adaptation

To answer our research questions defined in §1, we experimented with feature-based and instance-based DA algorithms. We focused on these techniques for two reasons. First, most dementia detection models to date are feature-based, owing in part to clinical interest in the characteristic language use by people with dementia. Second, the size of available dementia detection datasets (see Table 1) precludes the use of the same types of deep learning models that are common in many other NLP tasks. The prevalence of smaller scale, feature-based models suggests that these DA techniques hold greater immediate task relevancy.

AUGMENT. AUGMENT is a straightforward feature-based DA algorithm that has been shown to be effective on a wide range of datasets and tasks (Daumé III, 2007). It augments the feature space by making "source-only," "target-only," and "shared" copies of each feature, effectively tripling the feature set using the following formulation where $\phi^{\mathbf{s}}, \phi^{\mathbf{t}} : X \rightarrow \breve{X}$ represent mappings for the source and target data, respectively:

$$\phi^{\mathbf{s}}(\mathbf{x_i}) = \langle \mathbf{x_i}, \mathbf{0}, \mathbf{x_i} \rangle, \quad \phi^{\mathbf{t}}(\mathbf{x_i}) = \langle \mathbf{0}, \mathbf{x_i}, \mathbf{x_i} \rangle \quad (1)$$

In the formulation above, $\breve{X} = \mathbb{R}^{3F}$ is then the augmented version of the feature space $X = \mathbb{R}^F$. Empty vectors are filled with $\mathbf{0} = \langle 0, 0, ..., 0 \rangle \in \mathbb{R}^F$. The motivation behind AUGMENT is intuitive. If a column contains a feature that correlates with the class label in both the target and source data, the learning algorithm will weight the shared column more heavily and reduce the weight on the target-only and source-only feature copies, reducing their importance to the model. However, if a feature correlates with the class label only with target (or source) data, the learning algorithm will increase the weight of the target-only (or source-only) column and reduce the weight of the others. The onus is thus left to the model to learn feature importance with respect to the domains.

MULTIAUGMENT. We extend AUGMENT to accommodate multiple source domains following guidelines sketched out by Daumé III (2007), and refer to the technique as MULTIAUGMENT. As in the two-domain case, we expand the feature space, but this time to $\mathbb{R}^{(K+1)F}$ where K is the total number of domains. The cardinality $(k+1)F$ represents a distinct feature set F for each domain $k_i \in K$, plus the same shared feature space introduced previously. For our specific case we test this method with two source domains, creating the following mappings to transform from $\mathbb{R}^F$ to $\mathbb{R}^{4F}$:

$$\begin{aligned} \phi^{\mathbf{s_1}}(\mathbf{x_i}) &= \langle \mathbf{x_i}, \mathbf{0}, \mathbf{0}, \mathbf{x_i} \rangle, \\ \phi^{\mathbf{s_2}}(\mathbf{x_i}) &= \langle \mathbf{0}, \mathbf{x_i}, \mathbf{0}, \mathbf{x_i} \rangle \\ \phi^{\mathbf{t}}(\mathbf{x_i}) &= \langle \mathbf{0}, \mathbf{0}, \mathbf{x_i}, \mathbf{x_i} \rangle \end{aligned} \quad (2)$$

TRADABOOST. TRADABOOST is a supervised instance-based DA method (Dai et al., 2007) that extends the AdaBoost classification algorithm (Freund and Schapire, 1997) for transfer learning. The method is based on a "reverse boosting" principle, where the weights of poorly predictive source instances are decreased at each boosting iteration and the weights of target instances are simultaneously increased. The guiding intuition is that instances with large weights (including source instances that are more distributionally similar to the target domain instances) can then play a greater role in training the learning algorithm. We used the TRADABOOST implementation in Python's `adapt`

[2] `https://alphacephei.com/vosk/`

Group	# Features	Category
POS	12	*l*
CFG	12	*l*
Syntac. Complexity	16	*l*
NER	10	*l*
Vocab. Richness	6	*l*
SUBTL	1	*l*
Semantic	5	*s*
Acoustic	25	*a*

Table 2: Descriptive feature statistics. *Category* refers to the high-level categorization applied to features when performing experiments: *l*, *s*, and *a* are lexicosyntactic, semantic, and acoustic features, respectively.

package[3] to implement this technique.

3.4 Features

We experimented with lexicosyntactic, semantic, and acoustic features, summarized below. All features are calculated using the participant's utterances or speech segments. Descriptive statistics indicating the number of features belonging to each group, as well as the group's high-level categorization (used when labeling experimental conditions), are presented in Table 2.

Part-Of-Speech (POS) Tags. POS tags have proven useful for detecting dementia (Masrani, 2018), as well as primary progressive aphasia and two of its subtypes (Balagopalan et al., 2020b). We use the spaCy[4] core English POS tagger to capture the frequency of coarse-grained POS labels in a transcript using the Universal Dependencies tagset (Petrov et al., 2012). Frequency counts are normalized by the number of words in the transcript.

CFG Features. Context-Free Grammar (CFG) features count how often a phrase structure rule (e.g., $NP \rightarrow VP\ PP$ or $NP \rightarrow DT\ NP$) occurs in an utterance parse tree. These feature counts are then normalised by the total number of nodes in the parse tree. CFG features have previously demonstrated success for dementia detection (Masrani, 2018; Masrani et al., 2017). We extract parse trees using the Stanford parser (Qi et al., 2018), representing constituents using Penn Treebank constituent tags (Marcus et al., 1993).

Syntactic Complexity. Measures of syntactic complexity have proven effective for predicting dementia from speech (Masrani, 2018). We represent utterance complexity through a suite of features including parse tree depth, mean word length, mean sentence length, mean clause (noun or verb phrase) length, and number of clauses per sentence.

Named Entity Recognition (NER) Tags. Although NER features have *not* been studied in prior work, we suspected that they may be a useful and relatively domain-agnostic way to encode broad structural patterns, following the previous success of other more general intent-based features (Farzana and Parde, 2022). We extracted named entity labels using a spaCy[5] model trained on the OntoNotes 5 corpus. This model produces the fine-grained named entity types present in the OntoNotes tagset (Pradhan et al., 2007). We included a frequency feature for each NER type. NER frequency counts were normalized by the total number of entities mentioned in the transcript.

Vocabulary Richness Features. Existing research has shown that measures of vocabulary richness can be successfully leveraged to diagnose dementia (Masrani et al., 2017; Balagopalan et al., 2020a). We include a set of well-known lexical richness measures including type-token ratio (TTR), moving-average TTR (MATTR), mean segmental TTR (MSTTR), Maas index (Mass, 1972), the measure of textual lexical diversity (McCarthy, 2005, MTLD), and the hypergeometric distribution index (McCarthy and Jarvis, 2007, HD-D). We calculated each measure over the entire transcript using Python's `lexicalrichness` package.[6]

SUBTL Scores. SUBTL scores represent the frequency with which words are used in daily life (Brysbaert and New, 2009). They are derived from large corpora[7] of television and film subtitles spanning 50 million words. We treated tokens with the Penn Treebank POS tags PRP, PRP$, WP, and EX as stopwords and computed transcript-level SUBTL scores by averaging across all available word-level scores for the participant's speech.

Semantic Features. We measure semantic similarity between consecutive utterances by calculating the cosine similarity between the utterance

[3] https://adapt-python.github.io/adapt/generated/adapt.instance_based.TrAdaBoost.html

[4] https://spacy.io/usage/linguistic-features#pos-tagging

[5] https://spacy.io/api/annotation#named-entities

[6] https://pypi.org/project/lexicalrichness/

[7] http://www.lexique.org/

vectors and then recording the proportion of distances below three thresholds (0, 0.3, 0.5). We used averaged TF-IDF vectors to represent each utterance. We also recorded the minimum and average cosine distance between utterances.

Acoustic Features. Finally, prior work has found acoustic distinctions between subjects with and without dementia (Masrani et al., 2017). We chunked the participant's speech segments from each audiorecording using `Pydub`[8] prior to extracting acoustic features. We include prosody features (Dehak et al., 2007; Vásquez-Correa et al., 2018) from continuous speech based on duration (i.e., number of voiced segments per second and standard deviation of duration of unvoiced segments), extracted using the `DiSVoice`[9] tool.

4 Evaluation

4.1 Classification Settings

For our backbone classifier, we experimented[10] with support vector machine (SVM) and logistic regression (LR), implemented using `sklearn`.[11] For SVM, we used a polynomial kernel and held all other hyperparameters at their default settings except for the trade-off parameter C. For LR, we also held all hyperparameters at their default settings. We selected LR and SVM due to their documented success at dementia detection using one or more of our datasets (Farzana and Parde, 2020; Masrani et al., 2017). We tuned our models using K-fold stratified cross-validation on the training set, using the following values for the trade-off parameter C: {0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5, 1}.

We report the result for the parameter achieving the best performance, averaged across all five folds.[12] We used stratified cross-validation to produce the results reported in all results tables. We maintained the same ratio between the target classes in all folds and in the full dataset, and shuffled samples for cross-validation such that all samples from the same participant remained in the same fold. This was done to prevent overfitting due to data leakage stemming from the same participant being present in multiple folds.

[8] https://pypi.org/project/pydub/

[9] https://github.com/jcvasquezc/DisVoice

[10] Code for our experiments: https://github.com/treena908/Domain_Adaptive_Dementia_Detection

[11] https://scikit-learn.org/stable/

[12] We found that $C = 1$ resulted in the best performance across all folds.

4.2 Experimental Conditions

We compared each DA technique against three baseline models: a model jointly trained using samples from both the source and target data without applying any DA algorithms (JOINT), a model trained only on the target data (TARGET), and a model trained only on the source data (SOURCE). The training dataset(s) for our DA conditions varied depending on the technique being tested. AUGMENT and TRADABOOST were trained on data from a single source domain and the target domain, whereas MULTIAUGMENT was trained on data from two source domains and the target domain. All models, including the DA algorithms tested and our baseline models, were evaluated using the *target* domain test set.

We considered the following *source* → *target* adaptations: CCC → DB, DB → CCC, CCC → ADReSS, {ADRC, CCC} → DB, {ADRC, DB} → CCC, and {ADRC, CCC} → ADReSS. For each DA technique, we also considered several combinations of feature subsets (refer to Table 2 for categorizations): *l*, *l*+*s*, and *l*+*s*+*a*. MULTIAUGMENT only used *l* and *l*+*s* since ADRC does not provide speaker segmentation timestamps; thus, speech could not be extracted in the same way as other datasets, preventing use of acoustic features.

4.3 Results

We compared the conditions specified in §4.2 using accuracy and F_1, and report our experimental results in Tables 3, 4, and 5. Results are subdivided according to target domain, presenting results from conditions using DB, CCC, and ADReSS as the target domains, respectively.

We find that MULTIAUGMENT clearly outperforms the baseline techniques in most cases and usually outperforms the single-source DA algorithms when DB and ADReSS are the target domains, although the best-performing feature subsets vary. This trend is less clear when CCC is the target domain, with AUGMENT approaching or exceeding the performance of MULTIAUGMENT. When task-oriented data (DB or ADReSS) was used as the target, we observed that the percentage of source data in the training set was lower than that in the target data. As a result, we suspect that adding more conversational data (such as that found in ADRC) to the source (CCC) may promote improved performance when adapting to task-oriented target domains.

Model	LR Acc.	LR F1	SVM Acc.	SVM F1
SOURCE$_{L+S}$	0.45 (0.01)	0.02 (0.04)	0.45 (0.01)	0.01 (0.01)
TARGET$_{L+S}$	0.74 (0.05)	**0.77** (0.05)	0.67 (0.06)	0.61 (0.11)
JOINT$_{L+S}$	0.72 (0.06)	0.74 (0.06)	0.66 (0.04)	0.64 (0.06)
AUGMENT$_{L}$	0.72 (0.04)	0.75▼ (0.04)	**0.70** (0.04)	0.76▲ (0.04)
AUGMENT$_{L+S}$	0.73 (0.06)	0.75 (0.05)	**0.70** (0.05)	0.76▲ (0.04)
AUGMENT$_{L+S+A}$	0.72 (0.05)	0.74 (0.05)	0.69 (0.05)	0.74▲ (0.05)
TRADABOOST$_{L}$	0.66 (0.05)	0.68 (0.05)	0.66 (0.06)	0.68 (0.08)
TRADABOOST$_{L+S}$	0.60▼ (0.05)	0.64▼ (0.04)	0.63 (0.06)	0.64 (0.08)
TRADABOOST$_{L+S+A}$	0.55▼ (0.05)	0.55▼ (0.07)	0.65 (0.05)	0.68 (0.06)
MULTIAUGMENT$_{L}$	0.72▼ (0.06)	0.75▼ (0.06)	**0.70** (0.05)	0.77▲ (0.05)
MULTIAUGMENT$_{L+S}$	**0.75** (0.05)	0.76▼ (0.05)	**0.70** (0.05)	**0.77**▲ (0.05)

Table 3: Comparison of DA conditions when CCC is the source dataset and DB is the target dataset (standard deviation is reported inside parentheses). For MULTIAUGMENT conditions, ADRC and CCC are jointly used as the source dataset. Five-fold cross-validation is used in all cases with each fold having 40% source data (46.3% class *d*) and 60% target data (55.5% class *d*). MULTIAUGMENT has 46.3% source data (26.70% class *d*) and (53.70%) target data (55.2% class *d*). ▲: Significantly better than the corresponding LR or SVM TARGET baseline, with $p < 0.05$, using a paired *t*-test. ▼: Significantly worse than the corresponding TARGET baseline, using the same parameters.

Model	LR Acc.	LR F1	SVM Acc.	SVM F1
SOURCE$_{L+S}$	0.39 (0.07)	0.51 (0.05)	0.36 (0.02)	0.51 (0.02)
TARGET$_{L+S}$	0.83 (0.07)	0.72 (0.15)	0.80 (0.05)	0.63 (0.12)
JOINT$_{L+S}$	0.80 (0.10)	0.66 (0.18)	0.84 (0.08)	0.73 (0.16)
AUGMENT$_{L}$	0.85 (0.07)	**0.75** (0.15)	0.85 (0.06)	0.74 (0.15)
AUGMENT$_{L+S}$	0.84 (0.07)	0.73 (0.15)	0.84 (0.07)	0.73 (0.15)
AUGMENT$_{L+S+A}$	0.80 (0.06)	0.68 (0.15)	0.80 (0.05)	0.68 (0.11)
TRADABOOST$_{L}$	0.80 (0.05)	0.69 (0.13)	0.84 (0.07)	0.74 (0.17)
TRADABOOST$_{L+S}$	0.79 (0.05)	0.66 (0.14)	0.84 (0.06)	0.73 (0.15)
TRADABOOST$_{L+S+A}$	0.78 (0.06)	0.65 (0.14)	0.83 (0.08)	0.71 (0.17)
MULTIAUGMENT$_{L}$	**0.85** (0.07)	0.74 (0.15)	**0.86**▲ (0.07)	**0.75**▲ (0.16)
MULTIAUGMENT$_{L+S}$	0.84 (0.07)	0.73 (0.15)	0.85▲ (0.07)	0.74▲ (0.15)

Table 4: Comparison of DA conditions when DB is the source dataset and CCC is the target dataset (standard deviation is reported inside parentheses). For MULTIAUGMENT conditions, ADRC and DB are jointly used as the source dataset. Five-fold cross-validation is used in all cases with each fold having 70.3% source data (55.5% class *d*) and 29.7% target data (33.6% class *d*). MULTIAUGMENT has 73.1% source data (48.2% class *d*) and 56.9% target data (33.6% class *d*). ▲: Significantly better than the corresponding LR or SVM TARGET baseline, with $p < 0.05$, using a paired *t*-test. ▼: Significantly worse than the corresponding TARGET baseline, using the same parameters.

Both AUGMENT and MULTIAUGMENT outperform TRADABOOST, regardless of feature combination, across the board. We achieve maximum performance of F_1=0.77 on DB (using MULTIAUGMENT$_{L+S}$ with SVM), F_1=0.75 on CCC (using MULTIAUGMENT$_{L}$ with SVM), and F_1=0.77 on ADReSS (using AUGMENT$_{L+S}$, MULTIAUGMENT$_{L}$, and MULTIAUGMENT$_{L+S}$ with SVM). In Table 6, we report additional results from our highest-performing versions of each DA technique on the ADReSS test set (Luz et al., 2020). This facilitates straightforward comparison with external models by others who use this standardized test set. We find that AUGMENT$_{L+S}$ achieves similar results to those in Table 5.

5 Analysis

The results in Tables 3–6 clearly answer our first research question (Q1), demonstrating that DA can be used to exploit spoken language data pertaining to dementia detection in one domain to improve its detection in other domains. They also answer Q2, showing that DA offers performance improvements over jointly training on data from multiple domains. To answer Q3, we performed additional analyses to probe the contributions of feature subsets and class bias to overall performance.

5.1 Feature Analysis

To find correspondences between source and target domain features, we analyzed the features in the

Model	LR		SVM	
	Acc.	F1	Acc.	F1
SOURCE$_{L+S}$	0.52 (0.04)	0.06 (0.11)	0.51 0.04	0.03 (0.09)
TARGET$_{L+S}$	0.80 (0.13)	0.75 (0.22)	0.68 (0.12)	0.54 (0.25)
JOINT$_{L+S}$	0.69 (0.16)	0.64 (0.24)	0.59 (0.14)	0.47 (0.25)
AUGMENT$_{L}$	0.77 (0.12)	0.72 (0.20)	0.78▲ (0.10)	0.74▲ (0.17)
AUGMENT$_{L+S}$	0.74▼ (0.10)	0.68▼ (0.21)	**0.81**▲ (0.07)	**0.77**▲ (0.15)
AUGMENT$_{L+S+A}$	0.75 (0.06)	0.69 (0.14)	0.80▲ (0.15)	0.76▲ (0.22)
TRADABOOST$_{L}$	0.72▼ (0.14)	0.67▼ (0.21)	0.77 (0.13)	0.71▲ (0.21)
TRADABOOST$_{L+S}$	0.76 (0.13)	0.70 (0.21)	0.76 (0.13)	0.70▲ (0.21)
TRADABOOST$_{L+S+A}$	0.76 (0.12)	0.73 (0.17)	0.70 (0.10)	0.62 (0.19)
MULTIAUGMENT$_{L}$	0.80 (0.13)	0.75 (0.22)	0.80▲ (0.13)	**0.77**▲ (0.20)
MULTIAUGMENT$_{L+S}$	0.75 (0.14)	0.67 (0.29)	**0.81**▲ (0.14)	**0.77**▲ (0.21)

Table 5: Comparison of DA conditions when CCC is the source dataset and ADReSS train is the target dataset (standard deviation is reported inside parentheses). For MULTIAUGMENT conditions, ADRC and CCC are jointly used as the source dataset. Ten-fold cross-validation is used in all cases with each fold having 74.8% source data (33.6% class d) and 25.2% target data (50% class d). MULTIAUGMENT has 79.4% source data (26.70% class d) and 20.6% target data (50% class d). ▲: Significantly better than the corresponding LR or SVM TARGET baseline, with $p < 0.05$, using a paired t-test. ▼: Significantly worse than corresponding TARGET baseline, using the same parameters.

Model	C	LR		SVM	
		Acc.	F1	Acc.	F1
SOURCE$_{L+S}$	d c	0.51	0.00 0.68	0.51	0.00 0.68
TARGET$_{L+S}$	d c	0.72	0.65 0.76	0.70	0.53 0.77
JOINT$_{L+S}$	d c	0.68	0.62 0.73	0.72	0.70 0.75
AUGMENT$_{L+S}$	d c	**0.77**	0.70 **0.81**	**0.77**	**0.72** 0.80
MULTIAUGMENT$_{L+S}$	d c	0.74	0.68 0.79	0.74	0.68 0.79
TRADABOOST$_{L+S}$	d c	0.74	0.70 0.78	0.72	0.67 0.76

Table 6: Evaluation on the standardized ADReSS test set with per-class (C) F_1. CCC and ADReSS (train) are used as source and target data, respectively, when training with 46.1% source data (no class d) and 54.9% target data (55.5% class d). For MULTIAUGMENT, both CCC and ADRC are used as source (77.6% training data with 26.7% class d), with 32.4% target data (50% class d). We assessed statistical significance using McNemar's test, and found that no improvements were significantly different from the TARGET baseline.

shared column from AUGMENT$_{L+S+A}$ using LR and a DB $\rightarrow$ CCC domain adaptation mapping. We referred to these as *pivot features*. We computed the most important pivot features across source and target domain using $l1$-penalty with logistic regression.

We find that a subset of specific lexicosyntactic and acoustic pivot features, including the number of tokens, average phrase length, and standard deviation of the duration of unvoiced segments are highly positively correlated with the class labels in both the source and target domains. In contrast, the number of unique named entities, certain vocabulary richness and lexical frequency measures (MATTR and SUBTL score), and the number of voiced segments per second are highly negatively correlated with the class labels of both the source and target domains. Thus, these features offer particularly strong contributions to model performance across multiple domains.

5.2 Domain-Specific Class Bias

As shown in Table 1, our domains vary in their class balance. Class imbalances are especially common in low-resource healthcare tasks since it is often challenging to recruit subjects with the target condition. When the source and target domains have varying class distribution, they are biased towards different class labels. This can create conditions such that the learning algorithm is able to capitalize upon class bias rather than real properties of the data to increase perceived performance. For instance, when adapting from CCC $\rightarrow$ DB with the source dataset (CCC) having 33.6% instances belonging to class d and the target dataset (DB) having 55.5% instances belonging to class d, it is possible that the model trivially learns to predict class d with greater frequency, without learning real feature distinctions between the classes.

To investigate whether the improvements observed from DA in our case may simply be the product of domain-specific class biases, we con-

Domain	Class	cb1	cb2	cb3	cb4
CCC	d	72	57	42	28
	c	28	43	58	72
DB	d	72	57	42	28
	c	28	43	58	72

Table 7: Distribution of instances across domains (CCC=*source*; DB=*target*) and classes within training folds for four different class biases.

		LR		SVM	
Condition	**Model**	**Acc.**	**F1**	**F1**	**Acc.**
Equal	JOINT$_{L+S}$	0.62 (0.09)	0.59 (0.14)	0.58 (0.08)	0.37 (0.20)
	AUGMENT$_{L+S}$	**0.64** (0.08)	**0.63** (0.09)	**0.65** (0.09)	**0.64▲** (0.09)
Consistent	JOINT$_{L+S}$	0.65 (0.08)	0.59 (0.18)	0.60 (0.08)	0.41 (0.20)
	AUGMENT$_{L+S}$	**0.67** (0.05)	**0.64** (0.13)	**0.65** (0.08)	**0.63** (0.03)

Table 8: Domain-specific class bias results (standard deviation is reported inside parentheses). ▲: Significantly better than the corresponding LR or SVM JOINT baseline, with $p < 0.05$, using a paired *t*-test.

ducted an experiment analyzing performance of AUGMENT$_{L+S}$ (our best-performing model in terms of accuracy for the CC $\rightarrow$ DB mapping, shown in Table 3) and JOINT$_{L+S}$ across class-biased and unbiased subsets of the original dataset. In our *equal* condition, both domains had perfectly class-balanced data in each training fold. In our *consistent class bias* condition, training folds had the varying class biases shown in Table 7. Each class bias setting was evaluated using five-fold cross-validation, and then those results were averaged. We report the results from this experiment in Table 8. We find that AUGMENT still outperforms JOINT in both conditions, answering the second part of Q3 by empirically demonstrating that class bias does not account for the performance improvements resulting from domain adaptation.

6 Discussion and Conclusions

Our work reveals intriguing findings on the use of DA for dementia detection. First, we find that DA can be successfully leveraged to improve feature-based dementia detection performance. This is the most comprehensive study of feature-based DA for this task, and the first to consider instance-based DA. We find that feature-based DA outperforms instance-based DA, and that an approach allowing for multiple source domains (MULTIAUGMENT) holds promise in many cases. In general, F_1 score is similar across target datasets, ranging from 0.76 (CCC) to 0.77 (DB and ADReSS).

Our DA conditions also exhibit clear performance improvements over jointly training on the same data, offering further evidence to support the use of DA for this task. Finally, in follow-up studies on the importance of individual features and class biases in this setting, we find that pivot features pertaining to number of tokens, average phrase length, acoustic qualities, named entities, and measures of vocabulary richness and lexical frequency are particularly critical to strong performance. This suggests that these features may be particularly robust across domains. We also demonstrate that the performance of DA conditions relative to joint training is not due to domain-specific class bias, further strengthening our conclusions. In the future, we hope to conduct follow-up studies to further probe the limits and nuances of DA applied to this and other low-resource healthcare tasks.

7 Limitations

Our work is limited by several factors. First, we conduct our work primarily using popular, publicly available dementia detection datasets, all of which are in English. Thus, it is unclear whether our findings generalize to other languages, especially with richer morphology where different predictive patterns may emerge. Second, due to the emphasis on feature-based models in most dementia detection work, we study only feature-based and instance-based DA approaches. Neural DA approaches may yield different findings, although they are less relevant for many current dementia detection approaches. Finally, we only study two backbone classification algorithms in our experiments. These classifiers are among the most common in prior work with our selected datasets; however, it may be the case that with a wider scope, other classification algorithms may yield different results. Collectively, these limitations present intriguing avenues for follow-up work.

8 Ethical Considerations

This research was guided by a broad range of ethical considerations, taking into account factors associated with fairness, privacy, and intended use. Although many of these are described throughout the paper, we summarize those that we consider

most critical in this section. It is our hope that by building a holistic understanding of these factors, we develop improved perspective of the challenges associated with the study of low-resource healthcare problems and the positive broader impacts that they may create.

Data Privacy and Fairness. This research was approved by the Institutional Review Board at the University of Illinois Chicago. Access was granted for all datasets used in this research, and our use is governed by approved protocols unique to each dataset. DementiaBank, ADReSS, and the Carolina Conversations Collection are all publicly available following access request protocols specified by their governing organizations. We refer readers to the citations throughout this work if they are interested in obtaining access to this data. We are unable to share it directly, although we can share our processing scripts and other code to facilitate reproducibility of our work by others.

ADRC is a privately-held dataset collected in collaboration with clinical partners under a rigorous set of guidelines governed by a separate, approved Institutional Review Board protocol at the University of California San Diego. This dataset will eventually be released, following further manual review to ensure full de-identification, but it cannot yet be released at this time. The data is currently stored on a password-protected server under VPN protection. To maximize reproducibility of our work by others unable to immediately gain access to this dataset, we limit the use of this dataset to a small set of experimental conditions (specifically, those using MULTIAUGMENT).

Intended Use. Automated models for dementia detection from spoken language present potential benefits in real-world scenarios: they offer opportunity to expand healthcare access, minimize cost of care, and reduce caregiver burden. However, they may also pose risks if used in unintended ways. We consider intended use of the work reported here to extend to the following:

- People may use the technology developed in this work to study language differences between individuals with and without dementia, as a way of building further understanding of the condition.
- People may use the technology developed in this work to further their own research into low-resource NLP tasks, including those associated with this and other healthcare problems.
- People may use the technology developed in this work to build early warning systems to flag individuals about potential dementia symptoms, *provided that the technology is not misconstrued as an alternative to human care* in any way.

Any use outside of those listed above is considered an unintended use. To safeguard against unintended use of our work, we remind readers that dataset access must be granted through the approved channels by the creators of the respective datasets used in this work. This may include processes ranging from email request to full review and approval by local and external Institutional Review Boards. We reiterate our caution against using any findings from this paper to build systems that function as intended or perceived replacements for human medical care.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback, which was incorporated in the final version of this manuscript. We also thank Erin Sundermann for her and her team's role in creating the ADRC dataset, and Raeanne Moore, Alex Leow, and Tamar Gollan for their clinical insights regarding Alzheimer's disease and dementia. The creation of the ADRC dataset was funded in part by a seed grant from the University of California San Diego's Alzheimer's Disease Research Center. Shahla Farzana and Natalie Parde were also partially funded by the National Science Foundation under Grant No. 2125411. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Samad Amini, Boran Hao, Lifu Zhang, Mengting Song, Aman Gupta, Cody Karjadi, Vijaya B. Kolachalama, Rhoda Au, and Ioannis Ch. Paschalidis. 2022. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. *Alzheimer's & Dementia*, n/a(n/a).

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020a. To BERT or not to BERT: Comparing Speech and Language-Based

Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech 2020*, pages 2167–2171.

Aparna Balagopalan, Jekaterina Novikova, Matthew B A Mcdermott, Bret Nestor, Tristan Naumann, and Marzyeh Ghassemi. 2020b. Cross-Language Aphasia Detection using Optimal Transport Domain Adaptation. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 202–219. PMLR.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*.

Marc Brysbaert and Boris New. 2009. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41:977–90.

Hamidreza Chinaei, Leila Chan Currie, Andrew Danks, Hubert Lin, Tejas Mehta, and Frank Rudzicz. 2017. Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. *Computational Linguistics*, 43(2):377–406.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 193–200, New York, NY, USA. Association for Computing Machinery.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

BH Davis, C Pope, K Van Ravenstein, and W Dou. 2017. Three approaches to understanding verbal cues from older adults with diabetes. *The Internet Journal of Advanced Nursing Practice*, 16(1).

Najim Dehak, Pierre Dumouchel, and Patrick Kenny. 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103.

Flavio Di Palo and Natalie Parde. 2019. Enriching neural models with targeted features for dementia detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 302–308, Florence, Italy. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *CoRR*, abs/2002.06305.

Shahla Farzana and Natalie Parde. 2020. Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues. In *Proc. Interspeech 2020*, pages 2207–2211.

Shahla Farzana and Natalie Parde. 2022. Are interaction patterns helpful for task-agnostic dementia detection? an empirical exploration. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–182, Edinburgh, UK. Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Jeffrey M. Girard, Alexandria K. Vail, Einat Liebenthal, Katrina Brown, Can Misel Kilciksiz, Luciana Pennant, Elizabeth Liebson, Dost Öngür, Louis-Philippe Morency, and Justin T. Baker. 2022. Computational analysis of spoken language in acute psychosis and mania. *Schizophrenia Research*, 245:97–115. Computational Approaches to Understanding Psychosis.

Harold Goodglass and Edith Kaplan. 1972. *The assessment of aphasia and related disorders*. Lea & Febiger.

Sarah A. Graham, Ellen E. Lee, Dilip V. Jeste, Ryan Van Patten, Elizabeth W. Twamley, Camille Nebeker, Yasunori Yamada, Ho-Cheol Kim, and Colin A. Depp. 2020. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research*, 284:112732.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

Wouter M Kouw and Marco Loog. 2018. An introduction to domain adaptation and transfer learning. Technical report, Delft University of Technology.

Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA Open*, 3(2):146–150.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Saturnino Luz, Sofia De La Fuente Garcia, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. In *Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive impairment*, pages 35–42. European Language Resources Association (ELRA).

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *medRxiv*.

Brian Macwhinney. 2009. The CHILDES Project Part 1: The CHAT Transcription Format. Technical report, Carnegie Mellon University.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Vaden Masrani. 2018. Detecting dementia from written and spoken language. Master's thesis, University of British Columbia.

Vaden Masrani, Gabriel Murray, Thalia Shoshana Field, and Giuseppe Carenini. 2017. Domain adaptation for detecting mild cognitive impairment. In *Advances in Artificial Intelligence*, pages 248–259, Cham. Springer International Publishing.

Heinz-Dieter Mass. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Shamila Nasreen, Julian Hough, and Matthew Purver. 2021a. Rare-class dialogue act tagging for Alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–300, Singapore and Online. Association for Computational Linguistics.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021b. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, 3.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. 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. Association for Computational Linguistics.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2058–2065. AAAI Press.

J.C. Vásquez-Correa, J.R. Orozco-Arroyave, T. Bocklet, and E. Nöth. 2018. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of Communication Disorders*, 76:21–36.

Rui Xia, Jianfei Yu, Feng Xu, and Shumei Wang. 2014. Instance-based domain adaptation in nlp via in-target-domain logistic approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Ward Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3.2, 3.4, 4

☑ B1. Did you cite the creators of artifacts you used?
3.2, 3.4, 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3.2, 8

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.2, 4.1

C ☑ Did you run computational experiments?

4, 5

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4, 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3.2, 3.4, 4.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.