W3PHIAI-22 Data Hackallenge (Hackathon/Challenge)

Registration Required. Please sign up using the form here.

Background

Alzheimer's Disease is a debilitating condition with no known cure that affects every aspect of cognition, including language use. Over 50 million people are currently diagnosed with AD dementia, and this number is expected to triple by 2050 (Organization & others, 2017; Patterson, 2018; Prince et al., 2016). Diagnosis of AD dementia is time-consuming and challenging for patients and physicians alike, and currently relies on patient and caregiver reports, extensive neuropsychological examinations, and invasive imaging and diagnostic procedures. Individuals with unrecognized dementia suffer adverse outcomes ranging from anxiety over unexplained symptoms to family discord and catastrophic events (Boise et al., 1999; Bond et al., 2005; Stokes et al., 2015). While AD has no known cure, timely diagnosis can prevent or alleviate these adverse outcomes. Analysis of spoken language, which reflects cognitive status, suits this purpose; however, manually examining spoken language for markers of AD is not practical as it is time-consuming and requires extensive expertise. To this end, recent work has attempted to develop automated methods through which to diagnose dementia on the basis of language samples.

This work has been facilitated by the public availability of such samples from people with and without dementia. However a pervasive challenge in this work has been a difficulty in reproducing prior work and comparing results across studies on account of the use of different diagnosis-related subsets (e.g. probable vs. possible dementia), aggregation strategies (e.g. one vs. multiple transcripts per participant), performance metrics and cross-validation protocols. In this **hackachallenge,** we aim to address this problem directly by developing standardized analysis pipelines for two publicly-available datasets (the hackathon), providing a sound basis for meaningful comparative evaluation in the context of a shared task (the challenge).

Data

The Pittsburgh (Pitt) (Becker et al., 1994) and Wisconsin Longitudinal Study (WLS) (Herd et al., 2014) corpora are provided for this hackachallenge by the Dementia Bank consortium (https://dementia.talkbank.org/) in the same format as in Dementia Bank without any modifications specific to the hackachallenge. These two datasets consist of:

a) metadata containing demographic and diagnostic information and a wide range cognitive test results,
b) audio recordings of participants in the two studies describing a picture stimulus (Boston Diagnostic Aphasia Examination “Cookie Theft” stimulus), and
c) manual verbatim transcriptions of the audio recordings. Only the data provided to hackachallenge participants may be used - no external data will be allowed in order to establish a level playing field.

- **Dementia Bank Pitt Study Details:**
  A detailed description of this dataset is available in (Becker et al., 1994). In brief, the tests include a picture description task from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983), a widely-used diagnostic test for language abnormality detection. In this task, the participants are presented with a “Cookie Theft'” picture stimulus (see Figure 1) and asked to describe everything they see occurring in the picture. Participant responses were audio recorded and subsequently transcribed verbatim. Each participant was tested multiple times resulting in multiple transcripts per participant.

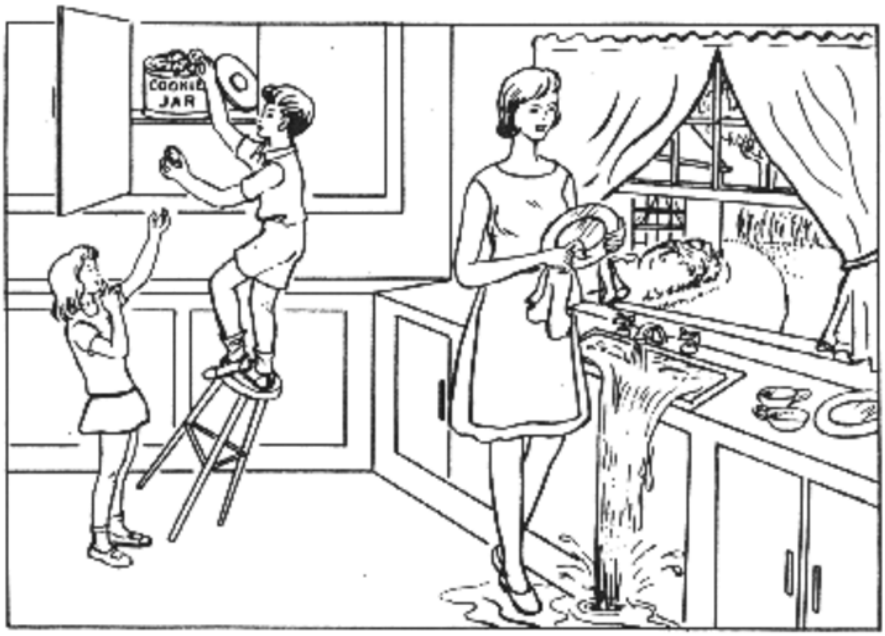

Figure 1: Cookie Theft stimuli

Further details of the contents of this corpus including media files and metadata with demographics and neuropsychological test scores are available here: https://dementia.talkbank.org/access/English/Pitt.html

Determining how to categorize these participants into “dementia” and “control” categories is not entirely straightforward. For example, typically studies that report results on this dataset rely on transcripts that are available for 169 subjects classified as having **possible** or **probable** AD dementia on the basis of clinical or pathological examination, and 99 patients classified as controls at the outset of the study. Of these 99 controls, 10 subsequently received a dementia-related diagnosis (according to the metadata in the data.xls file). According to the metadata, 7 of these 10 patients had a probable AD diagnosis and 3 had an indeterminate diagnostic status at baseline. Thus, the remaining 89 of the 99 patients represent the set of controls that are less likely to have had any underlying AD pathology present but undetected at baseline. In addition, only 148 of the dementia cases are assigned the category “probable AD” which indicates a higher level of diagnostic certainty for AD dementia.

Furthermore, there are 242 audio recordings/transcripts for the 99 controls and 257 recordings/transcripts for the 169 cases in this dataset. Many of the participants had multiple visits over the duration of this longitudinal study. The

number of visits for each participant is variable. This presents an additional challenge for analyzing these data, as participants' diagnostic status may change over time (as noted above) and one has to select a strategy for selecting recordings/transcripts or aggregating them in some way for feature engineering for predictive modeling. For example, for a given feature such as language model perplexity, the perplexity values across multiple transcripts per subject can be aggregated by taking the mean across all available samples, minimum or maximum value, only taking the first or the last available transcript, calculating a linear slope of the values over time or fitting a nonlinear function, among many other possible strategies.

- **Wisconsin Longitudinal Study Details:**
  The Wisconsin Longitudinal Study (WLS) (Herd et al., 2014) is a longitudinal study of a random sample of 10,317 graduates from Wisconsin high schools in 1957. The study also includes a randomly selected sibling of graduates, and spouses of graduates and siblings. WLS participants were interviewed up to six times across 60-years between 1957 and 2011. Beginning in 1993, during the fourth round of interviews, the WLS included cognitive evaluations. The "Cookie Theft" task was administered in the 2011 wave of the study (see Herd et al., 2014 for details).

  Further details of the contents of this corpus including media files and metadata with demographics and neuropsychological test scores are available here: https://dementia.talkbank.org/access/English/WLS.html

  All of the participants in the WLS study were considered to be cognitively healthy upon entry into the study. Some may have developed dementia in later years; however, the neurologic diagnostic information is not currently publicly available. Some investigators have used available cognitive test scores in order to identify participants whose cognitive test results are consistent with some form of cognitive impairment associated with dementia (e.g., category verbal fluency tests scores below age and education adjusted norms: see (Guo et al., 2021) for example). Prior work on verbal fluency performance in participants with AD established that "animal naming" category fluency scores less than 15 animal names in 60 seconds are 20 times more likely in a patient with dementia than in a healthy individual and were found to discriminate between these two groups with sensitivity of 0.88 and specificity of 0.96 (Duff-Canning et al., 2004). Using the metadata that is available for this corpus, It is also possible to filter

and/or separate participants based on presence or absence of self-reported memory and other cognitive complaints as well as self-reported mental health diagnoses.

All of the audio recordings of the “Cookie Theft” task for WLS participants have been transcribed by a professional transcription service as described in Guo et al. 2021 and these transcripts are made available as part of the download of the corpus. **Important Note:** Some of the transcripts have undergone additional detailed revision/correction by the Dementia Bank and linked with the audio at the utterance level. These transcripts are currently available in the folders “00” and “01” in the WLS distribution. The remainder of the transcripts are in their original uncorrected format and have not been linked with the audio files. The “linking” means that each utterance/sentence in the transcript has been timestamped with utterance start and end offsets from the beginning of the corresponding audio file. Transcripts that are not “linked” (those outside of folders 00 and 01) do not have this timing information.

Objectives

Develop an **analysis pipeline** for combining data from two separate datasets of audio recordings and transcriptions of picture descriptions performed as part of neuropsychological testing and with the corresponding metadata to **discriminate** between or **characterize** participants with dementia and healthy controls.

Ground Rules

In this hackachallenge, we invite groups and individual participants to compete with each other to determine who can develop an analytical pipeline that has the best categorization performance. Each group will be asked to **select** data samples from the provided data using criteria specified by the hackachallenge organizers based on available metadata (e.g., age range, sex, education range, diagnoses, cognitive scores, etc.). Each group will also be asked to **label** each data sample as one of two categories “positive” vs. “negative” using **one or more** of the criteria provided by the organizers described in the section **Criteria for defining categories**. Each team may determine their own cutoffs for the provided criteria, if they feel comfortable doing so, or use the suggested cutoffs provided by the organizers. Each group will compete by developing an analytical **method** (pipeline) of their choosing for discriminating between these

categories. Finally, each group will be asked to select an **evaluation** strategy of their choosing (e.g., leave-one-out, leave-two-out, k-fold cross validation, etc.).

A key feature of this challenge is that, in addition to the analytical pipeline, each group will be required to develop a **manifest** that provides the details of the group's experimental set up containing a list of data samples and their category labels, and details of the evaluation strategy (e.g., number of folds in k-fold cross-validation and which samples were used in each fold). The purpose of this manifest is to enable other groups to replicate the original experimental conditions and data so that they can apply their own analytical pipelines to the same problem to obtain results that are directly comparable across the groups. The details of the pipeline should NOT be included in the manifest and should be kept confidential by each group until the end of the hackachallenge at which point each group will be asked to provide a description of the methods that were used to obtain the results submitted to the competition.

We provide the pre-processing toolkit to allow participants to create customized and parameterized pre-processed transcripts while minimizing the variance of preprocessing. Also, we will provide an example manifest to illustrate what is expected from each group and a set of baseline results.

Process

The hackachallenge will proceed in two phases:

Phase 1 (pre-workshop):

- Download and become familiar with the datasets
- Pre-process data (audio and text) – e.g. remove/convert some of the tags, apply NLP and/or ASR, add new tags, etc.
- Define and create class labels (e.g. dementia vs. controls or dementia vs. mild cognitive impairment vs. controls)
- Develop and apply a custom pipeline
- Develop and publish a machine-readable manifest/makefile with details of data selection, pre-processing parameters, class definitions, evaluation strategy
- We will provide a fine-tuned BERT model results as the baseline

Phase 1 examples:

The competition will be performed in a round-robin fashion in which each group will produce two artifacts: an analytical method/pipeline (M) and a machine-readable manifest with any

necessary scripts (S) to implement the strategies stated in the manifest. For example, for a hackachallenge with three teams, Phase 1 will look as follows:

- Group A: Analytical Method M1 + manifest/scripts S1
- Group B: Analytical Method M2 + manifest/scripts S2
- Group C: Analytical Method M3 + manifest/scripts S3

Workshop

- Report the results from Phase 1, along with details for class label creation and preprocessing

Phase 2 (post-workshop):

In this phase each group will take the other groups' manifests and run their own analysis pipeline using the data selection, class definition, and evaluation strategy information provided in the other group's manifest to replicate the other group's experimental design so that the results can be directly compared.
Each group will report the results and compare them to the results published by other groups on the same manifest

Phase 2 examples

- Group A: S2 & S3 ->> evaluate on M1
- Group B: S1 & S3 ->> evaluate on M2
- Group C: S1 & S2 ->> evaluate on M3

Evaluation Criteria

Pipeline performance – ability of a group's analysis pipeline to outperform other group's pipelines on the same manifest in their ability to discriminate between classes. We encourage participants to use the following evaluation criteria:

a. Accuracy at flat rate
b. Accuracy at equal error rate
c. AUC: the area under the receiver operating characteristic curve, or ROC curve
d. Precision
e. Recall

f. F-measure (i.e., F_1 score)

Definitions of Evaluation Criteria

Given the following confusion matrix

		Predicted condition	
	Total population (P+N)	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True Negative (TN)

- Accuracy (ACC)
  - The proportion of the predicted labels for a sample must *exactly* match the corresponding pre-assigned label.
  - $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision
  - The percentage of classified positive samples are *actually* correct.
  - $\frac{TP}{TP+FN}$
- Recall
  - The percentage of actual positives was identified correctly.
  - Also called true positive rate (TPR).
  - $\frac{TP}{TP+FN}$
- False positive rate (FPR)
  - The percentage of false positives against all positive predictions
  - $\frac{FP}{TP+TN}$
- F_1score
  - The harmonic mean of precision and recall.
  - $\frac{2TP}{2TP+FP+FN}$
- ROC curve

- A graphical plot to illustrate the diagnostic ability of a binary classifier as its discriminate threshold varies as illustrated in Figure 2. Area under curve (AUC) represents the area under the entire ROC curve, which provides an aggregate measure of performance across all possible classification thresholds
- A measurement to optimize the trade-off between false positives and false negatives.

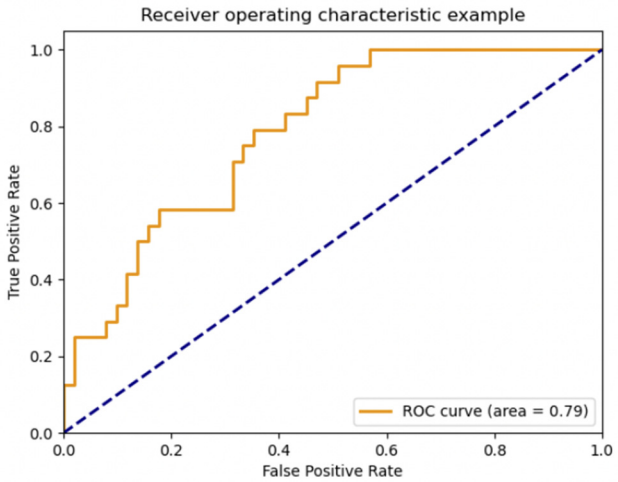

Figure 2: Example ROC curve. Source: [scikit-learn package website]

- Equal error rate (EER)
  - EER point is defined as the operating point where the false positive rate (FPR) and false negative rate (FNR) is equal
  - To calculate the EER, you need to first compute FPR, TPR, FNR and TNR, respectively.
  - ACU@EER is calculated based on the FPR and TPR
  - To calculate ACC@EER, you need to calculate prevalence threshold, which is defined as $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$, then the TPR@EER and TNR@EER are defined as the points/values that

Publication

The results of this challenge will be published in a journal that is TBD.

Criteria for Defining Categories

Due to differences in availability of diagnostic and neuropsychological assessments results in the two datasets, different criteria would need to be applied in order to define the "positive" and "negative" categories in the two datasets.

Dementia Bank Pitt Corpus: The metadata for the Pitt corpus is contained in a spreadsheet 'data.xls' and the definitions of the variables are provided in the Readme.txt file also available as part of the data download. The metadata contains the following variables most relevant to labeling individual data samples:

1. "dx1", "dx2", "dx3" - primary, secondary and tertiary diagnostic codes at baseline - the diagnoses that were given at the onset of the study. For example, a participant may have a diagnosis of "probable AD" (diagnostic code 100) and a secondary diagnosis of Parkinson's disease (diagnostic codes 420 or 430). Same for control participants (diagnostic code 800) - some controls may have secondary diagnoses of anxiety/depression (diagnostic code 730). Each team may decide how to use these codes in order to label the samples. For example, if the classification objective is to learn how to separate healthy controls from cases of AD dementia specifically, one could select all participants with code 800 as primary diagnosis at baseline and no secondary or tertiary diagnoses as the control group. It may also be important to ensure that none of the selected healthy controls transitioned to a diagnosis of mild cognitive impairment or probable AD at one of the subsequent visits. One could also select participants with a code 100 or 200 for the primary diagnosis.

   For a more challenging task, the teams may also choose participants with primary diagnostic codes of 720 and 740 that indicate mild cognitive impairment rather than (or in addition to) the probable AD diagnosis.

   An even more challenging objective would be to learn how to separate participants with probable AD from those with mild cognitive impairment.

2. Mini-Mental State Exam (MMSE) scores - The teams may choose to use MMSE scores instead or in addition to the diagnostic codes or any other criteria. MMSE scores above 24 are commonly found in cognitively healthy participants, whereas scores between 21 and 24 indicate possible Mild Cognitive Impairment (MCI), and scores of 21 or less suggest moderate dementia (Creavin et al., 2016; Dick et al., 1984).
3. Clinical Dementia Rating (https://knightadrc.wustl.edu/cdr/cdr.htm) - This (Cohen & Pakhomov, 2020) is a scale used in clinical settings to determine the stage of dementia. While the scale values may appear to be continuous and range between 0 and 3, the

scale is NOT continuous but rather ordinal - the value of 0 means “no dementia”, the value of 0.5 means “very mild dementia”, the value of 1 means “mild dementia”, the value of 2 means “moderate dementia” and the value of 3 means “severe dementia”

Dementia Bank WLS Corpus: As mentioned before, the WLS corpus consists of assessments of generally healthy participants at the onset of the study. Due to its relatively large size, this corpus may be a good source of healthy controls for training ML models; however, care needs to be taken in selecting samples that most likely belong to cognitively healthy individuals. There are several variables in the metadata provided with this corpus that could be used to find participants that are less suitable as controls.

1. “memory, 2011” -- this variable captures the participant’s self-report of how they would describe their ability to remember things during the past four weeks prior to the assessment.
1. “thinking, 2011” -- during the past four weeks, how would participant describe their ability to think and solve day to day problems
2. “stroke, 2011” -- has a doctor ever told the participant they had a stroke?
3. “mental illness, 2011” -- has the participant ever been diagnosed with a mental illness?

In addition to these self-reported variables, one may choose to use more objective scores obtained as part of cognitive testing in the two waves of the study for which the data are available - 2004 and 2011. The cognitive test of category verbal fluency has been shown to be very sensitive to the effects of dementia and has been recommended as a screening instrument for dementia for clinical use. The version of the category fluency test that is most frequently administered and for which there is most normative data available is the “animal” category test. In this test, the participant is asked to name all animals they can think of in one minute. The test is scored by counting how many correct animal names were produced excluding repetitions and hyponyms (e.g., dog and poodle count as one animal). In order to use verbal fluency scores from the WLS dataset, there are two variables that need to be involved for each of the waves - 2004 and 2011:

2004

1. “category fluency version, 2004” -- this field has two values: animals or food
2. “category fluency, # scored words produced, 2004” - this variable reflects the official verbal fluency score

2011

3. “category fluency version, 2011” -- this field has two values: animals or food

4. “category fluency, # scored words produced, 2011” - this variable reflects the official verbal fluency score

Prior work on verbal fluency performance in participants with AD established that animal fluency scores <15 are 20 times more likely in a patient with AD than in an healthy individual and were found to discriminate between these two groups with sensitivity of 0.88 and specificity of 0.96 (Canning et al., 2004). Recognizing the fact that verbal fluency performance does vary slightly by age and education(Marceaux et al., 2019; Tombaugh et al., 1999), a statistically determined age and education-adjusted thresholds of 16, 14, and 12 for participants in <60, 60–79, and >79 age ranges, respectively, may be used to identify low performers. We do not have normative data available for the food category; however, since the distributions of semantic verbal fluency scores on the “animal” category and “food” category tend to be very similar in this dataset, similar cutoffs could be used for the food category as for the animal category.

References

Becker, J., Boller, F., Lopez, O., Saxton, J., & McGonigle, K. (1994). The natural history of alzheimer’s disease. Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, *51*(6), 585–594.

Boise, L., Camicioli, R., Morgan, D. L., Rose, J. H., & Congleton, L. (1999). Diagnosing dementia: Perspectives of primary care physicians. *The Gerontologist*, *39*(4), 457–464. https://doi.org/10.1093/geront/39.4.457

Bond, J., Stave, C., Sganga, A., Vincenzino, O., O’connell, B., & Stanley, R. L. (2005). Inequalities in dementia care across Europe: Key findings of the Facing Dementia Survey. *International Journal of Clinical Practice*, *59*(s146), 8–14. https://doi.org/10.1111/j.1368-504X.2005.00480.x

Canning, S. D., Leach, L., Stuss, D., Ngo, L., & Black, S. (2004). Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology*, *62*(4), 556–562.

Cohen, T., & Pakhomov, S. (2020). A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1946–1957. https://doi.org/10.18653/v1/2020.acl-main.176

Creavin, S. T., Wisniewski, S., Noel-Storr, A. H., Trevelyan, C. M., Hampton, T., Rayment, D., Thom, V. M., Nash, K. J., Elhamoui, H., Milligan, R., & others. (2016). Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *Cochrane Database of Systematic Reviews*, *1*.

Dick, J. P., Guiloff, R. J., Stewart, A., Blackstock, J., Bielawska, C., Paul, E. A., & Marsden, C. D. (1984). Mini-mental state examination in neurological patients. *Journal of Neurology, Neurosurgery & Psychiatry*, *47*(5), 496–499. https://doi.org/10.1136/jnnp.47.5.496

Goodglass, H., & Kaplan, E. (1983). *Boston diagnostic aphasia examination booklet*. Lea & Febiger.

Guo, Y., Li, C., Roan, C., Pakhomov, S., & Cohen, T. (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*, 26. https://doi.org/10.3389/fcomp.2021.642517

Herd, P., Carr, D., & Roan, C. (2014). Cohort Profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology*, *43*(1), 34–41. https://doi.org/10.1093/ije/dys194

Marceaux, J. C., Prosje, M. A., McClure, L. A., Kana, B., Crowe, M., Kissela, B., Manly, J., Howard, G., Tam, J. W., Unverzagt, F. W., & others. (2019). Verbal fluency in a national sample: Telephone administration methods. *International Journal of Geriatric Psychiatry*, *34*(4), 578–587.

Organization, W. H. & others. (2017). *Global action plan on the public health response to dementia 2017–2025*.

Patterson, C. (2018). *World Alzheimer report 2018: The state of the art of dementia research:New frontiers.*

Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., & Karagiannidou, M. (2016). *World Alzheimer report 2016: Improving healthcare for people living with dementia: Coverage, quality and costs now and in the future*.

Stokes, L., Combes, H., & Stokes, G. (2015). The dementia diagnosis: A literature review of information, understanding, and attributions. *Psychogeriatrics : The Official Journal of the Japanese Psychogeriatric Society*, *15*(3), 218–225. https://doi.org/10.1111/psyg.12095

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, *14*(2), 167–177.