Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership

Chelsea Chandler[1], Peter W. Foltz[1,2], Jian Cheng[3], Jared C. Bernstein[3], Elizabeth P. Rosenfeld[3], Alex S. Cohen[4], Terje B. Holmlund[5], and **Brita Elvevåg[5,6]**
[1]University of Colorado Boulder, {chelsea.chandler, peter.foltz}@colorado.edu
[2]Pearson [3]Analytic Measures Inc.
[4]Louisiana State University
[5]University of Tromsø [6]Norwegian Centre for eHealth Research

Abstract

Verbal memory is affected by numerous clinical conditions and most neuropsychological and clinical examinations evaluate it. However, a bottleneck exists in such endeavors because traditional methods require expert human review, and usually only a couple of test versions exist, thus limiting the frequency of administration and clinical applications. The present study overcomes this bottleneck by automating the administration, transcription, analysis and scoring of story recall. A large group of healthy participants (n = 120) and patients with mental illness (n = 105) interacted with a mobile application that administered a wide range of assessments, including verbal memory. The resulting speech generated by participants when retelling stories from the memory task was transcribed using automatic speech recognition tools, which was compared with human transcriptions (overall word error rate = 21%). An assortment of surface-level and semantic language-based features were extracted from the verbal recalls. A final set of three features were used to both predict expert human ratings with a ridge regression model (r = 0.88) and to differentiate patients from healthy individuals with an ensemble of logistic regression classifiers (accuracy = 76%). This is the first 'outside of the laboratory' study to showcase the viability of the complete pipeline of automated assessment of verbal memory in naturalistic settings.

1 Introduction

Assessing human memory is one of the most important ways in which neurocognitive function is established. Memory is of central interest in numerous neurodevelopmental, neurodegenerative and neuropsychiatric conditions, as well as in brain injuries that affect cortical and subcortical brain systems (Baddeley and Wilson, 2002).

Given the importance of verbal memory, it is a core component of the globally employed Wechsler Memory Scale (Wechsler, 1997). The Logical Memory subtest requires the repetition of short stories that have been spoken by the examiner, both immediately and after a delay. Administering these assessments requires participants to be physically present with the examiner, who then gives scores manually by assigning points for key words or thematic units correctly recalled. The required time-consuming human review combined with the availability of only a couple of test versions limits their use and as such contributes to the bottleneck in the assessment of verbal memory. This unfortunately translates into infrequent assessments.

Automating certain aspects of such assessments holds promise of enabling more regular assessments as well as remote ones, which may be beneficial for monitoring treatment effectiveness and may also avert tragedy. Given that the verbal memory task is spoken, it is well-suited for automatization by leveraging recent advances in speech technology and machine learning. It has become possible to assess not just the words generated, but deeper measures of semantic understanding, which can be used to develop objective and sensitive metrics from the speech of patients with dementia (Fraser and Hirst, 2016; Yancheva and Rudzicz, 2016; Zhou et al., 2016), aphasia (Fraser et al., 2013), autism (Losh and Gordon, 2014; Prud'hommeaux et al., 2017; Goodkind et al., 2018), and mental illness (Elvevåg et al., 2007, 2010; Rosenstein et al., 2015; Bedi et al., 2015; Iter et al., 2018; Corcoran et al., 2018; Holmlund et al., 2019b). However, it is now time to move beyond simple proof of concept and translate such findings into viable clinical tools (Foltz et al., 2016). Indeed, since machine learning based approaches make it possible to mimic the actual assessment processes employed by expert humans,

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 137–147
Minneapolis, Minnesota, June 6, 2019.

the modeling and prediction of cognitive functions can be done by a machine in much the same manner as by humans. Thus, the entire pipeline can be automated from administration and transcription, to analysis and actual scoring of memory recall.

In the present research, we applied computational approaches to the speech generated from participants retelling stories from the verbal recall task in order to characterize the quality of their recall and determine the accuracy of this characterization. The approach developed natural language processing (NLP) measures that were designed to align with features related to verbal memory and story recall in order to best assess the data. This study focused on two computational tasks: 1) automatically assigning ratings to participants' retells based on how much of the content from the original story they remembered, and 2) performing a classification task to distinguish psychiatric patients from healthy participants. The study further examined how well these measures can be incorporated into a full analysis pipeline starting from data collection on a mobile platform outside of the traditional laboratory (thus in the real-world, perhaps noisy, environment), to automated speech recognition (ASR), and then to the conversion of the language to predictions of recall quality.

2 Related Work

NLP has been used in a range of clinical applications from detecting depression in twitter feeds (Coppersmith et al., 2015) to analyzing coherence in patient-clinician interactions (Elvevåg et al., 2007). In each case, text is reduced to a set of variables to relate to clinical measures of interest.

There are several classes of variables that can encode characteristics of texts. One class of measures are considered surface features of language. This includes counts of words, phrases, and words related to cognitive and affective processes (Pennebaker et al., 2015; Prud'hommeaux and Roark, 2011). A second class of measures examines structural features of language, such as parses of the syntactic structure, the probabilities that word pairs would likely occur together (e.g., n-grams), and the cohesion and coherence of a text. Finally, semantic features assess the meaning expressed in texts, such as choice of words as they relate to a specific topic, as well as encoding the underlying meaning of words, sentences, or whole passages. Such measures are often based on corpora that encode general knowledge of the world or a domain to measure meaning at a conceptual level rather than through the counting of direct overlap.

Previous studies have measured story recall by computing the distance between two pieces of text (Lehr et al., 2012, 2013). For example, a participant's retell can be compared against the original story to determine the amount of information retained. One approach to measuring this distance is computing a word alignment between the texts, which relies on participants using *exact* words and phrases to achieve a high memory score (Prud'hommeaux and Roark, 2011). A more robust approach is to measure the distance in a derived embedding space between two pieces of text. Latent Semantic Analysis (LSA) (Landauer et al., 1998) applies a singular value decomposition to a matrix of word-document co-occurrences in a large corpus. It then uses the cosine distance between representations which is able to account for semantic relationships in which a participant may make small changes in concepts such as "store" and "market". Recent studies (Dunn et al., 2002; Rosenstein et al., 2014) have used LSA to successfully model recall data from the Logical Memory subtest of the Wechsler Memory Scale to quantify the degradation of performance with increasing retrieval intervals.

More recently, word embedding models have been applied to assessing clinical discourse. Iter et al. (2018) modeled the coherence of patient discourse using LSA, word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). While LSA derives its semantic context from a bag-of-words across documents, the word2vec word embedding model derives its representation by considering the contexts in which each word appears by examining the window of words around each word. This window measures context, either taking into account the order of the words in that window or independent of word order. An advantage of the latter approach is that the method learns both semantic associations and syntactic word order.

3 Data

The present study was the result of data collection through a mobile application for the purpose of longitudinally tracking the mental state of psychiatric patients (Cohen et al., 2019; Holmlund et al., 2019a). The application is composed of a num-

ber of assessment tasks that engage participants in spoken and touch-based interactions in order to capture daily measures of cognition, affect and clinical state.

As part of the overall examination, participants' verbal memory was assessed. Stories were presented orally in a male voice and the participant was then immediately asked to retell the story with as many details as possible. After a delay of approximately one day, they were prompted to retell the same story. Each participant was presented with one new story per session and all stories were sampled across participants. There were a total of 24 different stories developed to be structurally similar to the Logical Memory subtest of the Wechsler Memory Scale-III (Wechsler, 1997). Multiple versions were created to enable frequent administration, as there exist only two test versions in the Wechsler Memory Scale which limits the frequency of administration and hence its clinical application. The stories were narrative in nature and ranged from 61 to 82 words in length. They each had two characters, a setting, an action that caused a problem, and a resolution. An example story is as follows:

> "On Monday morning, the woman woke up more tired than usual. When she walked downstairs to make herself a cup of coffee, she found her husband in the kitchen. She was surprised because he usually left an hour before she woke up. Her husband greeted her and reminded her that daylight savings time was over. Realizing the clocks were wrong, she happily ran upstairs and jumped back into bed."

Since this research concerned itself with evaluating the viability of leveraging speech technologies to automate a traditional verbal memory task, our focus was on usability engineering to ensure a robust design that could be implemented on a large scale, out of the controlled laboratory, and self-administered by the participant themselves. Therefore, the traditional matching of groups was not considered a priority, and nor is this feasible in machine learning studies that seek sample sizes in the thousands. Our participants comprised 105 stable patients with mental illness at a substance use treatment program and 120 undergraduate students at Louisiana State University presumed to be healthy (henceforth termed 'healthy participants'; see Holmlund et al., 2019a for details). This research program was approved by the relevant ethics committee (LSU Institutional Review Board #3618) and participants provided their informed written consent to this study. The 105 patients produced 750 retell responses, of which 575 were immediate retells and 175 were delayed retells. Each patient produced between 2 and 19 retells, with an average of 7.35 and standard deviation of 4.50. The 120 healthy participants produced 427 retell responses, of which 216 were immediate retells and 211 were delayed responses. Each produced between 2 and 15 retells with an average of 4.97 and standard deviation of 2.76. The scale of the collected data was impressive in size and quality given that an experimenter was not present during administration.

4 Human Rating of Story Recall

The audio of the memory recalls were transcribed by humans. Trained human raters read the transcriptions and assigned scores on the quality and amount of concepts and themes recalled, including characters, events, dates, descriptors, and feelings. The scores assigned were on a scale from 1 to 6, with 1 indicating no details were recalled, and 6 indicating all major and almost all minor concepts and themes were recalled. The responses were rated by three trained human raters with clinical experience. A subset (326) of the responses were rated by two independent raters in order to verify inter-rater reliability ($r = 0.92$). The high degree of agreement suggests that the rating rubric was reliable and thus appropriate for use in training a machine learning algorithm.

Over all the ratings, healthy participants generally received higher ratings for the amount of content recalled from the original story. For the immediate retells, they received an average rating of 4.31 ($SD = 1.38$) as compared to patients' average rating of 3.15 ($SD = 1.44$, $t = 9.5$, $p < .001$). The biggest differentiator between the two groups was in delayed retell (healthy participants average = 3.95, $SD = 1.45$; patients average = 2.24, $SD = 1.66$; $t = 9.8$, $p < .001$). Figure 1 shows that the average ratings assigned to patients on both the immediate and delayed retell were significantly lower than the average ratings assigned to healthy participants. The wide error bars indicate a large variability in the averages among both groups.

The two groups of participants differed both in

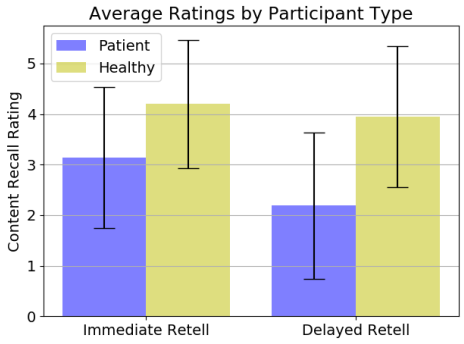

Figure 1: Average ratings by participant type. Error bars represent standard deviations of the samples.

the number of words typically spoken in a retell, and also in the relevance of their retell to the original story. While the histograms in figure 2 are somewhat biased since there was an uneven breakdown in the number of samples analyzed between the two groups, they do show that the peak of the distribution for patients is skewed more to the lower word counts than the peak for the healthy participants.

A noteworthy observation from the data is the amount of missing or silent responses. The tasks were self-administered by the participant outside of a traditional controlled setting, and there were several responses that were either silent or along the lines of "I don't remember". As expected, this type of missing data was more prevalent among patients than healthy participants, with 5% of the patient immediate retells and 19% of the patient delayed retells being less than 5 words or silent. While this is a constraint in live data collection in uncontrolled real world settings, it is a trait of realistic data that it will never be perfect and forced the creation of models capable of generating predictions on imperfect data. Instead of including silent responses (and thus allowing a classifier to learn that this is a trait common to patients), all silent responses were eliminated in order to create models that learn based on the language production, not the lack of any language.

5 Overview of Analysis Approach

There were four major components to this study. The first was feature engineering in order to determine a set of features that could be instantiated through computational NLP approaches and would assess important aspects of recall. We narrowed the large feature set down to only those most relevant to the constructs of story memory. Second, we built a regression model that could predict the ratings an expert human would assign to a story recall. Third, in order to show the predictive power of our data, we used the same features in a classification model to predict whether a participant was a patient or healthy participant. Fourth, in order to fully automate the pipeline, these analyses were completed on transcripts derived using ASR rather than the human transcriptions.

6 Feature Engineering

In designing NLP-based features to assess recall, it was critical to consider what aspects were most significant. A retell can be characterized by the amount of information recalled, the level of detail, changes in structure, as well as the quality of expressed language. Linguistic surface features provide indications of the overall amount of information recalled. Overuse of particular parts of speech, such as determiners, have been shown to provide indications of language ability, in that certain language constructions may indicate more sophisticated ability (Bedi et al., 2015). Retells, however, are affected by transformations of words within semantic memory (Kintsch, 1988). Indeed, surface features of a story (e.g., exact wording) are quickly lost in memory, but the gist is retained. Although a story may contain the word "market", a person may recall it as a "store". Thus, features that can account for semantics may be more effective at measuring the degree to which a memory has changed, with subtle effects of synonymy. Therefore, we investigated a variety of feature types ranging from linguistic surface features such as word counts to semantic features like cosine similarity between embedded representations.

The surface features included either raw or normalized counts of the number of tokens (word count), types (unique word count), n-grams (counts of word sequences of length n), or particular parts of speech. The surface features, while not the most sophisticated, nonetheless proved to be highly predictive. For instance, a simple count of the tokens informed how detailed the retell was. Whether the details aligned with the original story or not was revealed by the more advanced surface and semantic features. We further explored the use of specific parts of speech and ambiguous pronoun usage as Iter et al. (2018) concluded these are

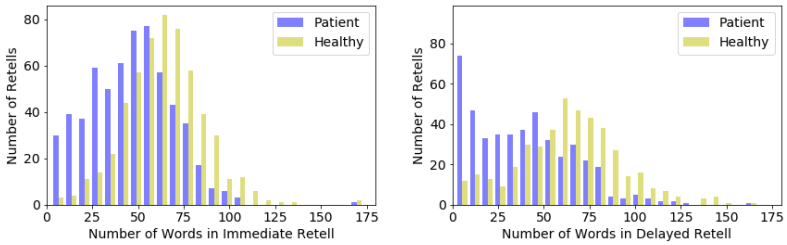

Figure 2: Immediate (left) and delayed (right) retell word count histograms by participant type.

traits of disordered speech. Since our data were composed of short responses that were fairly constrained in content, these features did not prove to be especially useful.

A step beyond raw counts is overlapping surface features, e.g. alignment, between the original story and the retell (Prud'hommeaux and Roark, 2012). The number of overlapping types between the original story and the retell measured how many concepts were remembered. For instance, if a retell stated that the event took place on "an afternoon" when it was actually "a rainy afternoon" in the original story, the type overlap can pick up on a missing detail. These counts offered a semantic relatedness indicator since recall of words from the original prompt was a good measure of memory, however, more interesting were metrics that could measure semantic similarity directly, somewhat independent of surface features.

Semantic features can be analyzed by using different types of embedded representations and metrics to score the distance between these representations. Word embeddings are widely employed to represent the semantic content as well as syntactic relationships of variable-length pieces of text. In this study, we tested pre-trained word embeddings, including word2vec and GloVe, and found that the pre-trained word2vec Google News corpus word embedding model (3 million 300-dimensional embeddings) produced results most correlated with our data.

Calculating the cosine distance between the average (both tf-idf weighted and unweighted) of the word embedding representations of two documents is a standard metric in NLP. We tested this in the current study, as well as the word mover's distance (WMD). Cosine distance was not as effective as WMD as it tends to smooth out the importance of individual words.

WMD is a good metric for analysing recall data as it captures word meaning and how semantically distant each word in a document is to its closest aligned word in another document. Thus, for verbal memory assessment, it provides a way to characterize how much semantic change there is from the original story to the recalled story. Put simply, WMD finds a mapping from each word in a document to its closest counterpart in the other and the distance is calculated as the sum of all Euclidean distances between matched words. Figure 3 illustrates the WMD calculation on document 1 (D1) and document 2 (D2) from a single source document (D0). Ignoring stop words, the model first finds a pairing between the most semantically similar words in the two documents. The arrows drawn between words in the documents represent a matching and are labeled with their distance contribution. WMD calculates a total distance as a function of all word pairings. D1 and D2 have an equal 'bag of words' distance of 0 from D0 as there are no overlapping content tokens, but semantically, D1 is much closer than D2. WMD is a more sophisticated method than cosine distance and has been shown to outperform it in many classification tasks (Kusner et al., 2015). For example, we compared the embedded representation for each participant's retell to the embedded representation for each original story using both the cosine metric and WMD, and overall the WMD metric correlated -0.82 with the human ratings while the cosine metric correlated -0.72.

A final feature considered was retell structure. Prior work has shown that language coherence can be useful clinically and predict risk of psychosis onset. To measure coherence, word embeddings are generated of n contiguous words in the retell and the semantic similarity to the embedded representation of the next n words is computed. Then the window is moved ahead by one word to make the next comparison, and then all the semantic similarities are averaged (Elvevåg et al., 2007). This approach provides a smoothed metric of the

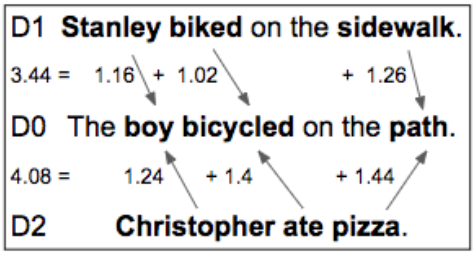

Figure 3: Example adapted from original paper by Kusner et al. (2015). The source sentence D0 and two query sentences, D1 and D2, aligned by words, and document distance computed by some function of total word pair distances.

cohesiveness of a retell, in that if the response is tangential or switches topics, it was assigned a lower overall coherence. In the present study, using a window size of four words on the retells correlated at 0.39 to the human ratings of the retells, indicating that better retells tended to be more coherent.

7 Human Rating Prediction Models

To fully automate the modeling of memory recall, a regression model was created that assigned a performance score to a story retell, treating immediate and delayed retells as the same task. Using a combination of univariate statistical tests and recursive feature elimination on the feature set, we identified the best combination of 3 features. They were not collinear and accounted for aspects of the rating task that align well with attributes that trained humans look for when rating recall. The features assessed the overall amount of content generated, the direct overlap of word types with the original story and the overall semantic change.

A ridge regression model was trained with a regularization parameter set to 0.01. We chose only three features to incorporate into the model in order to derive a system that is simple and interpretable. The three features used to generate ratings were the common types between the original story and retell (mean regression coefficient of 3.14), the word type count in the retell (mean regression coefficient of 2.47), and the word mover's distance between the original story and retell (mean regression coefficient of -2.71). 4 shows the correlations of each of the features to the rating given to the retell. The overall average correlation (Pearson r) with the human rating over 10-fold cross-validation through the data was 0.88. This average correlation of 0.88 of the model to the average human rating was in line with the 0.9 correlation between human raters. The implication is that automated assessment performs on par with humans, and additionally is an unbiased and convenient method. Success notwithstanding, it should be noted that the model performed poorly on responses that should have received low scores because key details of the original story were not recalled, but achieved a high word count, token overlap, and a reasonable word mover's distance. For instance, when a participant was prompted to retell the "balloon story" yet could not remember much, since prompted to talk about balloons, they were nonetheless able to ramble on about balloons, in essence 'fooling' the regression model.

8 Classification of Clinical Group Membership

The ability to automatically score recall is most definitely noteworthy, but the predictive power of the features was additionally demonstrated with a classification task which successfully identified the clinical group membership of the participant. Given that participants recalled each story twice, three classes of features were derived from the data: (i) how similar the initial retell was to the original story, (ii) how similar the delayed retell was to the original story, and (iii) how similar the initial retell was to the delayed retell.

As mentioned in the data section above, a goal of the current study was for the model to perform well in participants who were unable to complete both parts of the task. Therefore, an ensemble classifier was necessary to retain data for partial task completion. Each classifier made a classification based on features derived from a single session and the resulting subject-level classification was made from a combination of the individual session's prediction probabilities. This allowed silent or missing retells to be discarded yet still make predictions based on language data.

Prior applications of computational approaches in the cognitive health field have tended to perform classifications on a session-level (Prud'hommeaux and Roark, 2011; Rosenstein et al., 2014) rather than examining recall over multiple sessions. It was a goal of this research program to build a longitudinal model of behavior of an individual participant, so while the classifiers generated probability calculations at the session-level, all of these probabilities were aggregated over time and

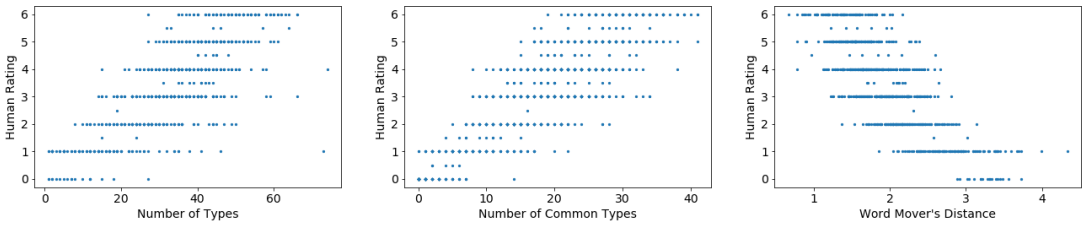

Figure 4: Scatter plots of our top features with human ratings. The number of common word types between the original story and the retell has a Pearson r correlation to average human ratings of 0.86, the number of word types in the retell has a Pearson r correlation of 0.82, and the word mover's distance between the original story and the retell has a Pearson r correlation of -0.82.

	Patient	Healthy
# immediate retells	575	216
# delayed retells	175	211
Average retells per participant	7.35, SD = 4.50	4.97, SD = 2.76
Range of retells per participant	[2,19]	[2,15]

Table 1: Breakdown of retell counts.

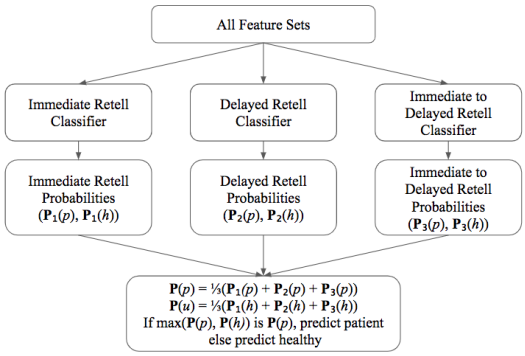

Figure 5: Ensemble Classifier Diagram.

tasks to make a final prediction at the subject-level. Leave-one-out cross-validation was performed across data from individual participants, training on all participants but one, and then subsequently testing on the one who was left out. Table 1 below contains a detailed breakdown of how many retells constituted profiles of the different groups, excluding any silent responses or responses with less than 5 words, which resulted in a disproportionate loss of delayed retells in patients.

The features used in the retell classifier were the number of unique types in the retell, the number of overlapping types between the original and retell, and the word mover's distance between the original and retell. Unsurprisingly, word mover's distance proved to be the most significant feature in the classifier. The delayed retell classifier was composed of the same features, but with calculations made on the delayed retell *in lieu* of the immediate retell. The last classifier, which focused on the change between the initial and delayed retell utilized two features: the number of common types between the immediate and delayed retell and the word mover's distance between the immediate and delayed retell.

The workflow for the ensemble classifier is shown in figure 5. The three classifiers were logistic regression classifiers optimized individually at a session-level. For instance, the retell classifier was trained on all retell features in the data, and predicted only on these inputs. Each classifier returned a tuple for each session of the probability that the session belonged to a healthy participant, $P_x(h)$, and to a patient, $P_x(p)$, where h represented the healthy class and p represented the patient class, x represented the classifier type, either $retell$, $reretell$, or $change$, and $P_x(h) + P_x(p) = 1$. The model then summed the probabilities coming from each classifier and normalized the summation to reach a final class membership probability.

$$P(p) = P_{retell}(p) + P_{reretell}(p) + P_{change}(p)$$
$$P(h) = P_{retell}(h) + P_{reretell}(h) + P_{change}(h)$$

The final prediction was a patient if $P(p) > P(h)$, otherwise it was a healthy participant. Put simply, the class with the largest overall probability was taken as the prediction.

The model correctly classified 78% of the patients and 74% of the healthy participants. Table 2 shows the confusion matrix from this classification model. The delayed retell classifier was the most accurate of the three as it was the biggest differentiator in performance between the two classes.

Patients misclassified as healthy had highly

	Predicted Healthy	Predicted Patient	Total
True Healthy	**89**	31	120
True Patient	23	**82**	105
Total	112	113	

Table 2: Confusion matrix of ensemble classification model. Model accuracy = 76%, precision = 73%, recall = 80%, F1 score = 76%.

rated retells that overlapped both semantically and at a word level with the original stories. Healthy participants misclassified as patients had multiple "I don't remember" or silent responses so their memory performance was characterised as poor.

9 Automated Speech Recognition

These results demonstrate that transcribed retells can be accurately characterized through computational methods. However, humans transcriptions in real-time are not practically viable. Therefore, to test how the methods would work in a fully automated pipeline, the same retells as generated by ASR were assessed. The audio files were run through two systems: (i) the latest Google Speech API (https://cloud.google.com/speech-to-text) which is a deep learning-based model trained on general English language, and (ii) a task-specific model that used a Deep Neural Network - Hidden Markov Model (Zhang et al., 2014) containing 5 hidden layers and 350 p-norm (p = 2) non-linearity neurons with a group size G = 10 per hidden layer, trained using Librispeech's (Panayotov et al., 2015) 960 hours of clean native (L1) reading data (Cheng, 2018). No speech data from the current dataset were used to train this acoustic model, but a 5-gram model based on the retells was used for the recognition.

Using Google's acoustic model, the average word error rate compared to human transcriptions across all patient retells was 26.51% and 16.38% across all healthy participant retells, totalling 20.90% on average. Using the task-specific model, the average patient word error rate was substantially less at 13.36% and the average healthy participant word error rate was 5.90% with an overall average of 10.79%. Some word errors were due to different word normalizations, for example "hashbrowns" versus "hash browns". Although transcriptions derived via ASR strayed somewhat from human transcriptions, the same ridge regression model described above, employing the same parameters, was then applied to the ASR-derived transcripts. As compared to the correlation r = 0.88 on a human transcription trained and tested regression model, the Google ASR trained and tested model achieved an r = 0.86, and the custom ASR trained and tested model achieved an r = 0.87. The change in performance on the classification ensemble model was similar; compared with an accuracy of 76% on the human model, both the Google ASR model and the custom ASR model achieved an accuracy of 74%. Thus, even with 10-20% word error rate, the model's predictive performance only lost a few percentage points, likely because it captures multiple aspects of the expressed language and so is highly robust to small errors if the overall sense is retained. The important implication is that audio collected from participants over mobile devices in realistic environments can be automatically transcribed with sufficient accuracy to provide useful predictions. Of note however, the nature of the current task and the fact that the retells had all been transcribed by humans who could screen for any potentially identifying information, ensured that there was zero risk of any identifying information being uploaded to the Google ASR system and thus critically maintained participant privacy. However, research that includes sensitive information (e.g., discussion of symptoms or things of a personal nature) must take additional measures to comply with relevant legislation and privacy protection rules.

10 Assessment Pipeline

This study - as illustrated in Figure 6 - demonstrates the solution to the bottleneck caused by time-consuming human review that is required in traditional settings and the resulting infrequent administration of verbal memory tests in current assessment practice. Our methodology enables the frequent and remote assessment of verbal memory and provides metrics of significant value in the new era of personalized medicine (Insel, 2017).

11 Conclusion

In conclusion, this study has overcome a classic bottleneck in traditional assessment practice and demonstrated that the promise of a truly personalized medicine approach to verbal memory assessment is realistic. The current study has validated the metrics on scores from expert human

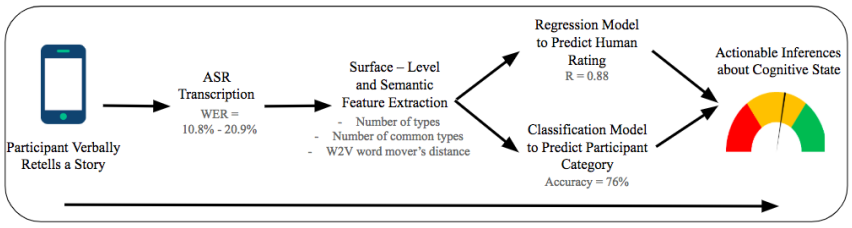

Figure 6: The complete pipeline of automated verbal memory assessment: It begins with a participant verbally retelling a story previously presented. Next, the retell is transcribed by an automatic speech recognition system. Once the speech is converted to text, various features are extracted, and sent to both a regression and classification model for ratings and categorization. Once complete, actionable inferences about cognitive state can be taken.

raters, and validated the actual assessment tool in terms of its functionality and usability. The design is demonstrably and sufficiently robust that this assessment tool is now ready to be applied within clinical settings to track patients longitudinally and inform clinicians accordingly. Future studies need to 'close the triage' by providing semi-immediate feedback from the assessment to the relevant entity. However, establishing the clinical and behavioral implications of such new metrics - such that they are calibrated correctly - remains an extremely complex empirical task which will necessitate the incorporation and modeling of multiple and dynamic data streams, as variables should not be interpreted in isolation when actionable clinical inferences are to be made.

12 Acknowledgements

This project was funded by grant 231395 from the Research Council of Norway awarded to Brita Elvevåg.

References

Alan Baddeley and Barbara A. Wilson. 2002. Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia*, (40(10)):1737–1743.

Gillinder Bedi, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández-Slezak, Mariano Sigman, Natalia B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, (1:15030).

Jian Cheng. 2018. Real-time scoring of an oral reading assessment on mobile devices. In *Proceedings Interspeech, Hyderabad, India, September 2–6*, pages 1621–1625.

Alex S. Cohen, Taylor Fedechko, Elana Schwartz, Thanh Le, Peter W. Foltz, Jared Bernstein, Jian Cheng, Elizabeth Rosenfeld, Terje B. Holmlund, and Brita Elvevåg. 2019. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *Journal of Abnormal Psychology*, (128):97–105.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd ACL Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Cheryl M. Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C. Javitt, Carrie E. Bearden, and Guillermo A. Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, (17(1)):67–75.

John C. Dunn, Osvaldo P. Almeida, Lee Barclay, Anna Waterreus, and Leon Flicker. 2002. Latent semantic analysis: A new method to measure prose recall. *Journal of Clinical and Experimental Neuropsychology*, (24(1)):26–35.

Brita Elvevåg, Peter W. Foltz, Mark Rosenstein, and Lynn DeLisi. 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*, (23):270–284.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, (93(1-3)):304–316.

Peter W. Foltz, Mark Rosenstein, and Brita Elvevåg. 2016. Detecting clinically significant events through automated language analysis: Quo imus? *npj Schizophrenia*, (2:15054).

Kathleen Fraser and Graeme Hirst. 2016. Detecting semantic changes in alzheimer's disease with vector space models. In *LREC*.

Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia.

In *4th Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54.

Adam Goodkind, Michelle Lee, Gary E. Martin, Molly Losh, and Klinton Bicknell. 2018. Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. In *Proceedings of the Society for Computation in Linguistics*, volume 1.

Terje B. Holmlund, Jian Cheng, Peter W. Foltz, Alex S. Cohen, and Brita Elvevåg. 2019b. Updating verbal fluency analysis for the 21st century: Applications for psychiatry. *Psychiatry Research*.

Terje B. Holmlund, Peter W. Foltz, Alex S. Cohen, H. D. Johansen, R. Sigurdsen, P. Fugelli, D. Bergsager, Jian Cheng, Jared Bernstein, Elizabeth Rosenfeld, and Brita Elvevåg. 2019a. Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: Practical challenges. *Psychological Assessment*, (31(3)):292–303.

Thomas R Insel. 2017. Digital phenotyping: Technology for a new science of behavior. In *JAMA*, 318(13), pages 1215–1216.

Dan Iter, Jong H. Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, pages 136–146.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review*, (95):163–182.

Matt Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, (25):259–284.

Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. *Proceedings of the 13th Annual Conference of the International Speech Communication Association.*, 2.

Maider Lehr, Izhak Shafran, Emily Prud'hommeaux, and Brian Roark. 2013. Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–220.

Molly Losh and Peter C. Gordon. 2014. Quantifying narrative ability in autism spectrum disorder: A computational linguistic analysis of narrative coherence. *Journal of Autism and Developmental Disorders*.

Thomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing*, pages 1532–1543.

Emily T. Prud'hommeaux and Brian Roark. 2011. Extraction of narrative recall patterns for neuropsychological assessment. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3021–3024.

Emily T. Prud'hommeaux and Brian Roark. 2012. Graph-based alignment of narratives for automated neurological assessment. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*.

Emily T. Prud'hommeaux, Jan van Santen, and Douglas Gliner. 2017. Vector space models for evaluating semantic fluency in autism. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 32–37.

Mark Rosenstein, Catherine Diaz-Asper, Peter W. Foltz, and Brita Elvevåg. 2014. A computational language approach to modeling prose recall in schizophrenia. *Cortex*, (55):148–166.

Mark Rosenstein, Peter W. Foltz, Lynn E DeLisi, and Brita Elvevåg. 2015. Language as a biomarker in those at high-risk for psychosis. *Schizophrenia Research*, (165):249–250.

David Wechsler. 1997. *Wechsler Memory Scale - Third Edition, WMS-III: Administration and scoring manual*. The Psychological Corporation.

Maria Yancheva and Frank Rudzicz. 2016. Vector-space topic models for detecting alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2337–2346.

Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. 2014. Improving deep neural network acoustic models using generalized maxout networks. In *ICASSP*, pages 215–219.

Luke Zhou, Kathleen Fraser, and Frank Rudzicz. 2016. Speech recognition in alzheimers disease and in its assessment. In *Interspeech 2016*, pages 1948–1952.