Using patterns of narrative recall for improved detection of mild cognitive impairment

Emily Prud'hommeaux
Oregon Health & Science Univ.
Portland, Oregon
emily@cslu.ogi.edu

Margaret Mitchell
University of Aberdeen
Aberdeen, Scotland, UK
m.mitchell@abdn.ac.uk

Brian Roark
Oregon Health & Science Univ.
Portland, Oregon
roark@cslu.ogi.edu

1 Introduction

The ability to identify dementia in its earliest stages is important not only for developing a treatment plan for patients but also for establishing support resources for families and caregivers. Unfortunately, Mild Cognitive Impairment (MCI), a frequent precursor to dementia, often goes undetected (Boise et al., 2004) in part because of an insufficient sensitivity of many common screening tests, such as the Mini Mental State Exam (Folstein et al., 1975), to subtle cognitive impairments (Shankle et al., 1996). As a result, a definitive diagnosis of MCI requires extensive examinations and interviews with the patient and caregiver.

Previous work (Shankle et al., 2005; Roark et al., In press) has shown that summary scores from simple linguistic memory tests such as the Wechsler Logical Memory test (Wechsler, 1997) and the CERAD word list recall tests (Welsh et al., 1994), can be used within a machine learning framework to improve the accuracy of detection of MCI. The present study focuses on leveraging information extracted from the Wechsler Logical Memory (WLM) subtests of the Wechsler Memory Scale to provide additional data that can be used in diagnosis. We investigate using the presence or absence of specific story elements in retellings as features within a machine learning framework to classify individuals into two groups: those with MCI and those without. Gathering this information places no burden on the examiner, as the story element identities are recorded during the standard administration the WLM. In addition, patients and caregivers are not required to complete any extra tasks or activities.

Our support vector machine classifier trained on story element scores achieves significantly higher accuracy than a classifier trained on the Logical Memory summary scores alone. In addition, combining CERAD word-list summary scores with story elements scores shows a significant improvement in accuracy over using CERAD summary scores alone. These results demonstrate the potential of using these previously unexplored but readily available features to enhance technology-assisted diagnosis of MCI.

2 Background

In the Logical Memory subtest of the Wechsler Memory Scale, a subject listens to a brief story and then retells the story to the examiner twice: once immediately upon hearing the story (Logical Memory I, LM-I), and a second time after a 30-minute delay (Logical Memory II, LM-II). Figure 1 shows the text of the Logical Memory narrative used in this study, with slashes indicating the boundaries between the brief phrases that constitute the story elements. During examination, the examiner notes which story elements the subject uses in each of his retellings. The subjects score is then calculated by counting the number of elements used in his retelling.

Note that the standard scoring procedure does not consider the identity of the story elements recalled. Rather, the summary score (i.e., the raw number of elements recalled) is the only score reported, even though the score sheet itself indicates which of the story elements were recalled.

Anna / Thompson / of South / Boston / employed / as a cook / in a school / cafeteria / reported / at the police / station / that she had been held up / on State Street / the night before / and robbed of / fifty-six dollars. / She had four / small children / the rent was due / and they hadn't eaten / for two days. / The police / touched by the woman's story / took up a collection / for her.

Figure 1: Text of Wechsler Logical Memory narrative, segmented into 25 story elements

3 Method

3.1 Subjects

Subjects in this study came from existing community cohort studies of brain aging at the NIA-funded Layton Aging & Alzheimer's Disease Center at Oregon Health & Science University. The Layton Center defines MCI in two ways: 1) via the Clinical Dementia Rating (CDR) scale (Morris, 1993), and 2) via a psychometrically-driven concept of degraded performance on a large set of neuropsychological tests. Following Shankle et al. (2005) and Roark et al. (In press), we defined our MCI and non-MCI groups based on the CDR. Since we are investigating the utility of different methods of deriving information from a particular neuropsychological test that might be used in the Layton Centers second definition of MCI, we used their first definition, the CDR scale, to provide an independent unconfounded reference objective for evaluation. The CDR has been shown to have high expert inter-annotator reliability (Morris et al., 1997) and, importantly, is assigned independently of the neuropsychological tests that we are investigating in this paper. We refer readers to the above cited papers for a full definition of the CDR.

We collected the original paper scoring sheets from just over 400 study participants, half of whom had received a CDR of 0.5, which corresponds to MCI, and the other half roughly age-matched individuals who have never had a CDR greater than 0. We chose the earliest available visit where the individuals had received the CDR of interest; i.e., for MCI subjects, the earliest visit where they received a CDR of 0.5, and for non-MCI subjects, their earliest visit.

We then manually entered the per-item results of the Wechsler Logical Memory test (both immediate and delayed) from these paper scoring sheets and reconciled the newly compiled results with what was in the database. Several subjects could not be included in this study due to mismatches between the data collected and the scores that should have been found for that session typically related to a failure to retrieve the correct trial scoring sheet from the files leaving 201 subjects with CDR 0 and 192 subjects with CDR 0.5. For all of these subjects, we have fully audited and validated per-item results for both immediate and delayed retellings of the Wechsler Logical Memory test. There were no significant between-groups differences in age or years of education. Details are presented in Table 1.

3.2 Features and Classification

As previously noted, the score sheets contain information not normally reported when scoring the Wechsler Memory Scale, namely, the identities of the recalled story elements. Thus, for each subject, we were able to assemble a feature vector composed of 52 features: one for each story element in LM-I, one for each element in LM-II, and summary scores for LM-I and LM-II. Each story element feature was assigned a binary value of 1 if the story element was recalled and 0 otherwise. Summary scores ranged from 0 (none of the 25 elements recalled) and 25 (all 25 elements recalled).

We used LibSVM (Chang and Lin, 2001), as implemented within the Waikato Environment for Knowledge Analysis (Weka) API (Hall et al., 2009), to train support vector machine (SVM) classifiers, using a second-order polynomial kernel and default parameter settings. Summary scores were scaled in both the training and testing data to range between 0 and 1, according to the minimum and maximum of the scores in the training data.

Measure	CDR = 0	CDR = 0.5
Age	81.2	79.7
Years education	14.5	14.5
Gender	78 M, 123 F	80 M, 112 F

Table 1: Demographic information.

Feature set	AUC	s.d.
LM summary scores	0.711	0.0260
LM story elements	0.827	0.0211
LM summary scores + story elements	0.827	0.0210
CERAD	0.836	0.0205
CERAD + LM summary scores	0.837	0.0205
CERAD + LM story elements	0.851	0.0197
CERAD + 7 chi-square-selected informative LM elements	0.885	0.0192

Table 2: Classification performance.

3.3 Evaluation

The performance of the SVM classifiers was evaluated using leave-one-out validation. In this validation method, each subject is tested against an SVM trained on all of the other subjects. The SVM per-subject scores can be used to evaluate the classifier quality according to one of the most commonly used classification evaluation methods: the Receiver Operating Characteristic (ROC) (Egan, 1975). The ROC plots the false positive rate of a classifier against the true positive rate. The area under the resulting curve (AUC) is the measure typically reported for accuracy. A random classifier would have an AUC of 0.5 (i.e., the area under the line from (0,0) to (1,1)), while a perfect classifier would have an AUC of 1.0. We use the Wilcoxon-Mann-Whitney statistic (Hanley and McNeil, 1982) to calculate the AUC.

In the final trial, we performed attribute selection to reduce the feature space of the set of story elements by ranking those features according to their chi-square statistic. Feature selection was performed separately on each training set to avoid introducing bias from the testing example. We trained and tested the SVM with the two CERAD scores and the top N story element features, from N=1 to N=50. We report here the accuracy for the top seven story elements (N=7), which yielded the highest AUC measure.

4 Results

To provide a baseline, we tested the SVM using a feature set consisting of the two LM summary scores alone as features. Subsequent trials used the following sets of features: all story elements, and all story elements together with the summary scores. Classification performance for these three features sets, is reported in rows 1-3 of Table 2.

We observe a dramatic increase in classification accuracy over the baseline by using the identities of the individual story elements as features. Including the summary scores together with the story elements did not improve performance, which suggests that the SVM is able to learn information about the summary scores from the element scores.

It has previously been shown that the CERAD word-list recall scores are also good predictors of MCI (Shankle et al., 2005). Since these scores are available for our pool of subjects, we now compare, in rows 4-7 of Table 2, the classification power of those scores with that of the Logical Memory summary scores and story elements scores.

The CERAD scores alone yield higher classification accuracy than the LM summary scores and slightly higher accuracy than the LM story element scores. However, including the LM story element scores with the CERAD scores in the SVM improves classification performance significantly over both of these feature sets individually. Furthermore, including only a subset of LM story elements, selected according to their predictive significance as measured by the chi-square statistic, improves accuracy dramatically, to 0.885.

5 Discussion and Future Work

The significant improvement in classification using the CERAD scores together with an informative subset of seven of the story element scores suggests that certain story elements may be more difficult to recall for subjects with MCI. Although we were careful to perform feature selection on the training data only, the same seven story element features were always

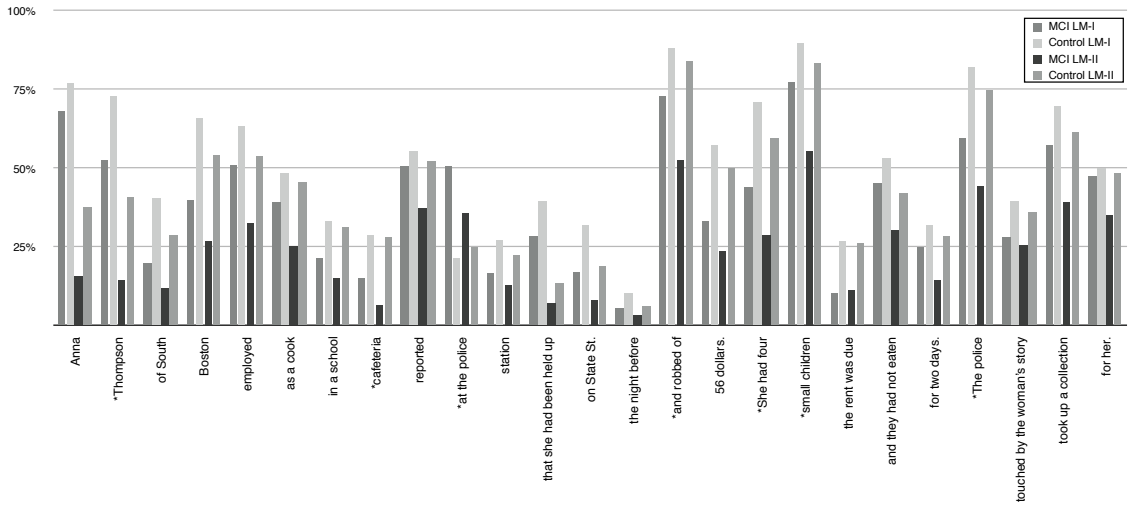

Figure 2: Percent of MCI and control subjects recalling each story element.

selected as the most informative. Figure 2 shows the percentage of subjects who recalled each of the story elements. The seven elements chosen with our feature selection method are denoted with an asterisk.

We observe that primacy and recency effects for both diagnostic groups are not as marked in narrative recall scenarios as they are typically reported to be in word-list recall scenarios (Shankle et al., 2005). The two most commonly recalled elements for both diagnostic groups, *small children* and *was robbed of*, fall very near the middle of the story. These frequently recalled elements are crucial plot points in the narrative, while the more rarely recalled items, such as*on State Street* and *the night before*, are minor details.

These two frequently recalled elements number among the seven most informative elements. We also see, however, that another of the most informative elements is *Thompson*, which is both early in the story and an incidental detail. Previous work has shown that event details with more structural and causal importance are more likely to be recalled in the unimpaired adult population (Johnson, 1970; Trabasso et al., 1984). Our future work will focus on determining how typical patterns of recall in unimpaired adults differ from those that are important for identification of MCI. In addition, we will explore using natural language processing techniques to automatically extract story elements from transcripts of LM narrative retellings. Other data that are not recorded during LM scoring but can be extracted from transcripts, such as element ordering and amount of relevant content, will also be investigated as potential features for the SVM classifier.

6 Conclusions

In this paper, we show that diagnostic classification for MCI can be significantly improved with the inclusion of Wechsler Logical Memory story elements. This data is already noted in the score sheet but is not considered in the standard scoring procedure and thus provides a readily available but previously untapped resource for improving the reliability of technology-based diagnosis of MCI.

References

Linda Boise, Margaret Neal, and Jeffrey Kaye. 2004. Dementia assessment in primary care: Results from a study in three managed care systems. *Journal of Gerontology*, 59(6).

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines.

James Egan. 1975. *Signal Detection Theory and ROC Analysis*. Academic Press.

M. Folstein, S. Folstein, and P. McHugh. 1975. Mini-mental state - a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12:189–198.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten.

2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

James Hanley and Barbara McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

R. E. Johnson. 1970. Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior*, 9:2–20.

J. Morris, C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L. Thal, and P. Woodbury. 1997. "clinical dementia rating training and reliability in multicenter studies: The alzheimer's disease cooperative study experience. *Neurology*, 48(6):1508–1510.

John Morris. 1993. The clinical dementia rating (CDR): Current version and scoring rules. *Neurology*, 43:2412–2414.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristina Hollingshead, and Jeffrey Kaye. In press. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing*.

W. Shankle, P. Datta, M. Dillencourt, and M. Pazzani. 1996. "improving dementia screening tests with machine learning methods. *Alzheimer's Research*, 2:95–99.

William R. Shankle, A. Kimball Romney, Junko Hara, Dennis Fortier, Malcolm B. Dick, James M. Chen, Timothy Chan, and Xijiang Sun. 2005. Methods to improve the detection of mild cognitive impairment. *Proceedings of the National Academy of Sciences*, 102(13):4919–4924.

T. Trabasso, T. Secco, and P. van den Broek. 1984. Causal cohesion and story coherence. In *Learning and comprehension of text*, page 83111. Erlbaum.

David Wechsler. 1997. *Wechsler Memory Scale - Third Edition Manual*. The Psychological Corporation.

K. Welsh, N. Butters, R. Mohs, D. Beekly, S. Edland, and G. Fillenbaum. 1994. The consortium to establish a registry for Alzheimer's disease (CERAD) part V. A normative study of the neuropsychological batte ry. *Neurology*, 44(4):609–614.