Graph-Based Word Alignment for Clinical Language Evaluation

Emily Prud'hommeaux*
Rochester Institute of Technology

Brian Roark**
Google, Inc.

Among the more recent applications for natural language processing algorithms has been the analysis of spoken language data for diagnostic and remedial purposes, fueled by the demand for simple, objective, and unobtrusive screening tools for neurological disorders such as dementia. The automated analysis of narrative retellings in particular shows potential as a component of such a screening tool since the ability to produce accurate and meaningful narratives is noticeably impaired in individuals with dementia and its frequent precursor, mild cognitive impairment, as well as other neurodegenerative and neurodevelopmental disorders. In this article, we present a method for extracting narrative recall scores automatically and highly accurately from a word-level alignment between a retelling and the source narrative. We propose improvements to existing machine translation-based systems for word alignment, including a novel method of word alignment relying on random walks on a graph that achieves alignment accuracy superior to that of standard expectation maximization-based techniques for word alignment in a fraction of the time required for expectation maximization. In addition, the narrative recall score features extracted from these high-quality word alignments yield diagnostic classification accuracy comparable to that achieved using manually assigned scores and significantly higher than that achieved with summary-level text similarity metrics used in other areas of NLP. These methods can be trivially adapted to spontaneous language samples elicited with non-linguistic stimuli, thereby demonstrating the flexibility and generalizability of these methods.

1. Introduction

Interest in applying natural language processing (NLP) technology to medical information has increased in recent years. Much of this work has been focused on information retrieval and extraction from clinical notes, electronic medical records, and biomedical academic literature, but there has been some work in directly analyzing the spoken language of individuals elicited during the administration of diagnostic instruments in clinical settings. Analyzing spoken language data can reveal information not only

* Rochester Institute of Technology, College of Liberal Arts, 92 Lomb Memorial Dr., Rochester, NY 14623. E-mail: emilypx@rit.edu.
** Google, Inc., 1001 SW Fifth Avenue, Suite 1100, Portland OR 97204. E-mail: roarkbr@gmail.com.

Submission received: 30 December 2013; revised submission received: 21 January 2015; accepted for publication: 4 May 2015.

doi:10.1162/COLI_a_00232

about impairments in language but also about a patient's neurological status with respect to other cognitive processes such as memory and executive function, which are often impaired in individuals with neurodevelopmental disorders, such as autism and language impairment, and neurodegenerative conditions, particularly dementia.

Many widely used instruments for diagnosing certain neurological disorders include a task in which the person must produce an uninterrupted stream of spontaneous spoken language in response to a stimulus. A person might be asked, for instance, to retell a brief narrative or to describe the events depicted in a drawing. Much of the previous work in applying NLP techniques to such clinically elicited spoken language data has relied on parsing and language modeling to enable the automatic extraction of linguistic features, such as syntactic complexity and measures of vocabulary use and diversity, which can then be used as markers for various neurological impairments (Solorio and Liu 2008; Gabani et al. 2009; Roark et al. 2011; de la Rosa et al. 2013; Fraser et al. 2014). In this article, we instead use NLP techniques to analyze the content, rather than the linguistic characteristics, of weakly structured spoken language data elicited using neuropsychological assessment instruments. We will show that the content of such spoken responses contains information that can be used for accurate screening for neurodegenerative disorders.

The features we explore are grounded in the idea that individuals recalling the same narrative are likely to use the same sorts of words and semantic concepts. In other words, a retelling of a narrative will be faithful to the source narrative and similar to other retellings. This similarity can be measured with techniques such as latent semantic analysis (LSA) cosine distance or the summary-level statistics that are widely used in evaluation of machine translation or automatic summarization, such as BLEU, Meteor, or ROUGE. Perhaps not surprisingly, however, previous work in using this type of spoken language data suggests that people with neurological impairments tend to include irrelevant or off-topic information and to exclude important pieces of information, or story elements, in their retellings that are usually included by neurotypical individuals (Hier, Hagenlocker, and Shindler 1985; Ulatowska et al. 1988; Chenery and Murdoch 1994; Chapman et al. 1995; Ehrlich, Obler, and Clark 1997; Vuorinen, Laine, and Rinne 2000; Creamer and Schmitter-Edgecombe 2010). Thus, it is often not the quantity of correctly recalled information but the quality of that information that reveals the most about a person's diagnostic status. Summary statistics like LSA cosine distance and BLEU, which are measures of the overall degree of similarity between two texts, fail to capture these sorts of patterns. The work discussed here is an attempt to reveal these patterns and to leverage them for diagnostic classification of individuals with neurodegenerative conditions, including mild cognitive impairment and dementia of the Alzheimer's type.

Our method for extracting the elements used in a retelling of a narrative relies on establishing a word alignment between a retelling and a source narrative. Given the correspondences between the words used in a retelling and the words used in the source narrative, we can determine with relative ease the identities of the story elements of the source narrative that were used in the retelling. These word alignments are much like those used to build machine translation models. The amount of data required to generate accurate word alignment models for machine translation, however, far exceeds the amount of monolingual source-to-retelling parallel data available to train word alignment models for our task. We therefore combine several approaches for producing reliable word alignments that exploit the peculiarities of our training data, including an entirely novel alignment approach relying on random walks on graphs.

Table 1
Abbreviations used in this article.

AER	alignment error rate
AUC	area under the (receiver operating characteristic) curve
BDAE	Boston Diagnostic Aphasia Exam
CDR	Clinical Dementia Rating
MCI	mild cognitive impairment
MMSE	Mini-Mental State Exam
WLM	Wechsler Logical Memory narrative recall subtest

In this article, we demonstrate that this approach to word alignment is as accurate as and more efficient than standard hidden Markov model (HMM)-based alignment (derived using the Berkeley aligner [Liang, Taskar, and Klein 2006]) for this particular data. In addition, we show that the presence or absence of specific story elements in a narrative retelling, extracted automatically from these task-specific word alignments, predicts diagnostic group membership more reliably than not only other dementia screening tools but also the lexical and semantic overlap measures widely used in NLP to evaluate pairwise language sample similarity. Finally, we apply our techniques to a picture description task that lacks an existing scoring mechanism, highlighting the generalizability and adaptability of these techniques.

The importance of accurate screening tools for neurodegenerative disorders cannot be overstated given the increased prevalence of these disorders currently being observed worldwide. In the industrialized world, for the first time in recorded history, the population over 60 years of age outnumbers the population under 15 years of age, and it is expected to be double that of children by 2050 (United Nations 2002). As the elderly population grows and as researchers find new ways to slow or halt the progression of dementia, the demand for objective, simple, and noninvasive screening tools for dementia and related disorders will grow. Although we will not discuss the application of our methods to the narratives of children, the need for simple screening protocols for neurodevelopmental disorders such as autism and language impairment is equally urgent. The results presented here indicate that the path toward these goals might include automated spoken language analysis.

2. Background

2.1 Mild Cognitive Impairment

Because of the variety of intact cognitive functions required to generate a narrative, the inability to coherently produce or recall a narrative is associated with many different disorders, including not only neurodegenerative conditions related to dementia, but also autism (Tager-Flusberg 1995; Diehl, Bennetto, and Young 2006), language impairment (Norbury and Bishop 2003; Bishop and Donlan 2005), attention deficit disorder (Tannock, Purvis, and Schachar 1993), and schizophrenia (Lysaker et al. 2003). The bulk of the research presented here, however, focuses on the utility of a particular narrative recall task, the Wechsler Logical Memory subtest of the Wechsler Memory Scale (Wechsler 1997), for diagnosing mild cognitive impairment (MCI). (This and other abbreviations are listed in Table 1.)

MCI is the stage of cognitive decline between the sort of decline expected in typical aging and the decline associated with dementia or Alzheimer's disease (Petersen et al. 1999; Ritchie and Touchon 2000; Petersen 2011). MCI is characterized by subtle deficits in functions of memory and cognition that are clinically significant but do not prevent carrying out the activities of daily life. This intermediary phase of decline has been identified and named numerous times: mild cognitive decline, mild neurocognitive decline, very mild dementia, isolated memory impairment, questionable dementia, and incipient dementia. Although there continues to be disagreement about the diagnostic validity of the designation (Ritchie and Touchon 2000; Ritchie, Artero, and Touchon 2001), a number of recent studies have found evidence that seniors with some subtypes of MCI are significantly more likely to develop dementia than the population as a whole (Busse et al. 2006; Manly et al. 2008; Plassman et al. 2008). Early detection can benefit both patients and researchers investigating treatments for halting or slowing the progression of dementia, but identifying MCI can be problematic, as most dementia screening instruments, such as the Mini-Mental State Exam (MMSE) (Folstein, Folstein, and McHugh 1975), lack sufficient sensitivity to the very subtle cognitive deficits that characterize the disorder (Morris et al. 2001; Ravaglia et al. 2005; Hoops et al. 2009). Diagnosis of MCI currently requires both a lengthy neuropsychological evaluation of the patient and an interview with a family member or close associate, both of which should be repeated at regular intervals in order to have a baseline for future comparison. One goal of the work presented here is to determine whether an analysis of spoken language responses to a narrative recall task, the Wechsler Logical Memory subtest, can be used as a more efficient and less intrusive screening tool for MCI.

2.2 Wechsler Logical Memory Subtest

In the Wechsler Logical Memory (WLM) narrative recall subtest of the Wechsler Memory Scale, the individual listens to a brief narrative and must verbally retell the narrative to the examiner once immediately upon hearing the story and again after a delay of 20 to 30 minutes. The examiner scores each retelling according to how many story elements the patient uses in the retelling. The standard scoring procedure, described in more detail in Section 3.2, results in a single summary score for each retelling, immediate and delayed, corresponding to the total number of story elements recalled in that retelling.

The Anna Thompson narrative, shown in Figure 1 (later in this article), has been used as the primary WLM narrative for over 70 years and has been found to be sensitive to dementia and related conditions, particularly in combination with tests of verbal fluency and memory. Multiple studies have demonstrated a significant difference in performance on the WLM between individuals with MCI and typically aging controls under the standard scoring procedure (Storandt and Hill 1989; Petersen et al. 1999; Wang and Zhou 2002; Nordlund et al. 2005). Further studies have shown that performance on the WLM can help predict whether MCI will progress into Alzheimer's disease (Morris et al. 2001; Artero et al. 2003; Tierney et al. 2005). The WLM can also serve as a cognitive indicator of physiological characteristics associated with Alzheimer's disease. WLM scores in the impaired range are associated with the presence of changes in Pittsburgh compound B and cerebrospinal fluid amyloid beta protein, two biomarkers of Alzheimer's disease (Galvin et al. 2010). Poor performance on the WLM and other narrative memory tests has also been strongly correlated with increased density

of Alzheimer related lesions detected in postmortem neuropathological studies, even in the absence of previously reported or detected dementia (Schmitt et al. 2000; Bennett et al. 2006; Price et al. 2009).

We note that clinicians do not use the WLM as a diagnostic test by itself for MCI or any other type of dementia. The WLM summary score is just one of a large number of instrumentally derived scores of memory and cognitive function that, in combination with one another and with a clinician's expert observations and examination, can indicate the presence of a dementia, aphasia, or other neurological disorder.

2.3 Previous Work

Much of the previous work in applying automated analysis of unannotated transcripts of narratives for diagnostic purposes has focused not on evaluating properties specific to narratives but rather on using narratives as a data source from which to extract speech and language features. Solorio and Liu (2008) were able to distinguish the narratives of a small set of children with specific language impairment (SLI) from those of typically developing children using perplexity scores derived from part-of-speech language models. In a follow-up study on a larger group of children, Gabani et al. (2009) again used part-of-speech language models in an attempt to characterize the agrammaticality that is associated with language impairment. Two part-of-speech language models were trained for that experiment: one on the language of children with SLI and one on the language of typically developing children. The perplexity of each child's utterances was calculated according to each of the models. In addition, the authors extracted a number of other structural linguistic features including mean length of utterance, total words used in the narrative, and measures of accurate subject–verb agreement. These scores collectively performed well in distinguishing children with language impairment, achieving an F1 measure of just over 70% when used within a support vector machine (SVM) for classification. In a continuation of this work, de la Rosa et al. (2013) explored complex language-model-based lexical and syntactic features to more accurately characterize the language used in narratives by children with language impairment.

Roark et al. (2011) extracted a subset of the features used by Gabani et al. (2009), along with a much larger set of language complexity features derived from syntactic parse trees for utterances from narratives produced by elderly individuals for the diagnosis of MCI. These features included simple measures, such as words per clause, and more complex measures of tree depth, embedding, and branching, such as Frazier and Yngve scores. Selecting a subset of these features for classification with an SVM yielded a classification accuracy of 0.73, as measured by the area under the receiver operating characteristic curve (AUC). A similar approach was followed by Fraser et al. (2014) to distinguish different types of primary progressive aphasia, a group of subtypes of dementia distinct from Alzheimer's disease and MCI, in a small group of elderly individuals. The authors considered almost 60 linguistic features, including some of those explored by Roark et al. (2011) as well as numerous others relating to part-of-speech frequencies and ratios. Using a variety of classifiers and feature combinations for three different two-way classification tasks, the authors achieved classification accuracies ranging between 0.71 and 1.0.

An alternative to analyzing narratives in terms of syntactic and lexical features is to evaluate the content of the narrative retellings themselves in terms of their fidelity to the source narrative. Hakkani-Tur, Vergyri, and Tur (2010) developed a method of

automatically evaluating an audio recording of a picture description task, in which the patient looks at a picture and narrates the events occurring in the picture, similar to the task we will be analyzing in Section 8. After using automatic speech recognition (ASR) to transcribe the recording, the authors measured unigram overlap between the ASR output transcript and a predefined list of key semantic concepts. This unigram overlap measure correlated highly with manually assigned counts of these semantic concepts. The authors did not investigate whether the scores, derived either manually or automatically, were associated with any particular diagnostic group or disorder.

Dunn et al. (2002) were among the first to apply automated methods specifically to scoring the WLM subtest and determining the relationship between these scores and measures of cognitive function. The authors used Latent Semantic Analysis (LSA) to measure the semantic distance from a retelling to the source narrative. The LSA scores correlated very highly with the scores assigned by examiners under the standard scoring guidelines and with independent measures of cognitive functioning. In subsequent work comparing individuals with and without an English-speaking background (Lautenschlager et al. 2006), the authors proposed that LSA-based scoring of the WLM as a cognitive measure is less biased against people with different linguistic and cultural backgrounds than other widely used cognitive measures. This work demonstrates not only that accurate automated scoring of narrative recall tasks is possible but also that the objectivity offered by automated measures has specific benefits for tests like the WLM, which are often administered by practitioners working in a community setting and serving a diverse population. We will compare the utility of this approach with our alignment-based approach subsequently in the article.

More recently, Lehr et al. (2013) used a supervised method for scoring the responses to the WLM, transcribed both manually and via ASR, using conditional random fields. This technique resulted in slightly higher scoring and classification accuracy than the unsupervised method described here. An unsupervised variant of their algorithm, which relied on the methods described in this article to provide training data to the conditional random field, yielded about half of the scoring gains and nearly all of the classification gains of what we report here. A hybrid method that used the methods in this article to derive features was the best performing system in that paper. Hence the methods described here are important components to that approach. We also note, however, that the supervised classifier-based approach to scoring retellings requires a significant amount of hand-labeled training data, thus rendering the technique impractical for application to a new narrative or to any picture description task. The importance of this distinction will become clear in Section 8, in which the approach outlined here is applied to a new data set lacking an existing scoring mechanism or a linguistic reference against which the responses can be scored.

In this article, we will be discussing the application of our methods to manually generated transcripts of retellings and picture descriptions produced by adults with and without neurodegenerative disorders. We note, however, that the same techniques have been applied to narratives transcribed using ASR output (Lehr et al. 2012, 2013) with little degradation in accuracy, given sufficient adaptation of the acoustic and language models to the WLM retelling domain. In addition, we have applied alignment-based scoring to the narratives of children with neurodevelopmental disorders, including autism and language impairment (Prud'hommeaux and Rouhizadeh 2012), with similarly strong diagnostic classification accuracy, further demonstrating the applicability of these methods to a variety of input formats, elicitation techniques, and diagnostic goals.

3. Data

3.1 Experimental Participants

The participants for this study were drawn from an ongoing study of brain aging at the Layton Aging and Alzheimer's Disease Center at the Oregon Health and Science University. Seventy-two of these participants had received a diagnosis of MCI, and 163 individuals served as typically aging controls. Demographic information about the experimental participants is shown in Table 2. There were no significant differences in age and years of education between the two groups. The Layton Center data included retellings for individuals who were not eligible for the present study because of their age or diagnosis. Transcriptions of 48 retellings produced by these ineligible participants were used to train and tune the word alignment model but were not used to evaluate the word alignment, scoring, or classification accuracy.

We diagnose MCI using the Clinical Dementia Rating (CDR) scale (Morris 1993), following earlier work on MCI (Petersen et al. 1999; Morris et al. 2001), as well as the work of Shankle et al. (2005) and Roark et al. (2011), who have previously attempted diagnostic classification using neuropsychological instrument subtest responses. The CDR is a numerical dementia staging scale that indicates the presence of dementia and its level of severity. The CDR score is derived from measures of cognitive function in six domains: Memory; Orientation; Judgment and Problem Solving; Community Affairs; Home and Hobbies; and Personal Care. These measures are determined during an extensive semi-structured interview with the patient and a close family member or caregiver. A CDR of 0 indicates the absence of dementia, and a CDR of 0.5 corresponds to a diagnosis of MCI (Ritchie and Touchon 2000). This measure has high expert inter-rater reliability (Morris 1993) and is assigned without any information derived from the WLM subtest.

3.2 Wechsler Logical Memory

The WLM test, discussed in detail in Section 2.2, is a subtest of the Wechsler Memory Scale (Wechsler 1997), a neuropsychological instrument used to evaluate memory function in adults. Under standard administration of the WLM, the examiner reads a brief narrative to the participant, excerpts of which are shown in Figure 1. The participant then retells the narrative to the examiner twice: once immediately upon hearing the narrative and a second time after 20 to 30 minutes. Two retellings from one of the participants in our study are shown in Figures 2 and 3. (There are currently two narrative

Table 2
Layton Center participant demographic data. Neither age nor years of education were significantly different between groups.

		Age (years)		Education (years)	
Diagnosis	*n*	**Mean**	**Std**	**Mean**	**Std**
MCI	72	88.7	6.0	14.9	2.6
Non-MCI	163	87.3	4.6	15.1	2.5

Anna / Thompson / of South / Boston / employed / as a cook / in a school / cafeteria / reported / at the police / station / that she had been [...] 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
Excerpts of WLM narrative with slashes indicating the boundaries between story elements. Twenty-two of the 25 story elements are shown here.

Ann Taylor worked in Boston as a cook. And she was robbed of sixty-seven dollars. Is that right? And she had four children and reported at the some kind of station. The fellow was sympathetic and made a collection for her so that she can feed the children.

Figure 2
WLM retelling by a participant before MCI diagnosis (score = 12).

She was robbed. And she had a couple children to feed. She had no food for them. And people made a collection for her and to pay for her, for the food for the children.

Figure 3
WLM retelling by same participant as in Figure 2 after MCI diagnosis (score = 5).

retelling subtests that can be administered as part of the Wechsler Memory Scale, but the Anna Thompson narrative used in the present study is the more widely used and has appeared in every version of the Wechsler Memory Scale with only minor modifications since the instrument was first released 70 years ago.)

Following the published scoring guidelines, the examiner scores the participant's response by counting how many of the 25 story elements are recalled in the retelling without regard to their ordering or relative importance in the story. We refer to this as the **summary score**. The boundaries between story elements are indicated with slashes in Figure 1. The retelling in Figure 2, produced by a participant without MCI, received a summary score of 12 for the 12 story elements recalled: *Anna, Boston, employed, as a cook, and robbed of, she had four, small children, reported, station, touched by the woman's story, took up a collection,* and *for her*. The retelling in Figure 3, produced by the same participant after receiving a diagnosis of MCI two years later, earns a summary score of 5 for the 5 elements recalled: *robbed, children, had not eaten, touched by the woman's story*, and *took up a collection*. Note that some of the story elements in these retellings were not recalled verbatim. The scoresheet provided with the exam indicates the lexical substitutions and degree of paraphrasing that are permitted, such as *Ann* or *Annie* for *Anna*, or any indication that the story evoked sympathy for *touched by the woman's story*. Although the scoring guidelines have an air of arbitrariness in that paraphrasing is only sometimes permitted, they do allow the test to be scored with high inter-rater reliability (Mitchell 1987).

Recall that each participant produces two retellings for the WLM: an immediate retelling and a delayed retelling. Each participant's two retellings were transcribed at the utterance level. The transcripts were downcased, and all pause-fillers, incomplete words, and punctuation were removed. The transcribed retellings were scored manually according to the published scoring guidelines, as described earlier in this section.

4. Diagnostic Classification Framework

4.1 Classifier

The goal of the work presented here is to demonstrate the utility of a variety of features derived from the WLM retellings for diagnostic classification of individuals with MCI. To perform this classification, we use LibSVM (Chang and Lin 2011), as implemented within the Waikato Environment for Knowledge Analysis (Weka) API (Hall et al. 2009), to train SVM classifiers, using a radial basis function kernel and default parameter settings.

We evaluate classification via receiver operating characteristic (ROC) curves, which have long been widely used to evaluate diagnostic tests (Zweig and Campbell 1993; Faraggi and Reiser 2002; Fan, Upadhye, and Worster 2006) and are also increasingly used in machine learning to evaluate classifiers in ranking scenarios (Cortes, Mohri, and Rastogi 2007; Ridgway et al. 2014). Analysis of ROC curves allows for classifier evaluation without selecting a specific, potentially arbitrary, operating point. To use standard clinical terminology, ROC curves track the tradeoff between **sensitivity** and **specificity**. Sensitivity (true positive rate) is what is commonly called **recall** in computational linguistics and related fields—that is, the percentage of items in the positive class that were correctly classified as positives. Specificity (true negative rate) is the percentage of items in the negative class that were correctly classified as negatives, which is equal to one minus the false positive rate. If the threshold is set so that nothing scores above threshold, the sensitivity (true positive rate, recall) is 0.0 and specificity (true negative rate) is 1.0. If the threshold is set so that everything scores above threshold, sensitivity is 1.0 and specificity is 0.0. As we sweep across intervening threshold settings, the ROC curve plots sensitivity versus one minus specificity, true positive rate versus false positive rate, providing insight into the precision/recall tradeoff at all possible operating points. Each point (tp, fp) in the curve has the true positive rate as the first dimension and false positive rate as the second dimension. Hence each curve starts at the origin (0, 0), the point corresponding to a threshold where nothing scores above threshold, and ends at (1, 1), the point where everything scores above threshold.

ROC curves can be characterized by the area underneath them ("area under curve" or AUC). A perfect classifier, with all positive items ranked above all negative items, has an ROC curve that starts at point (0, 0), goes straight up to (1, 0)—the point where true positive is 1.0 and false positive is 0.0 (since it is a perfect classifier)—before continuing straight over to the final point (1, 1). The area under this curve is 1.0, hence a perfect classifier has an AUC of 1.0. A random classifier, whose ROC curve is a straight diagonal line from the origin to (1, 1), has an AUC of 0.5. The AUC is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example, and is, in fact, equivalent to the Wilcoxon-Mann-Whitney statistic (Hanley and McNeil 1982). This statistic allows for classifier comparison without the need of pre-specifying arbitrary thresholds. For tasks like clinical screening, different tradeoffs between sensitivity and specificity may apply, depending on the scenario. See Fan, Upadhye, and Worster (2006) for a useful discussion of clinical use of ROC curves and the AUC score. In that paper, the authors note that there are multiple scales for interpreting the value of AUC, but that a rule-of-thumb is that AUC ≤ 0.75 is generally not clinically useful. For the present article, however, AUC mainly provides us the means for evaluating the relative quality of different classifiers.

One key issue for this sort of analysis is the estimation of the AUC for a particular classifier. Leave-pair-out cross-validation—proposed by Cortes, Mohri, and Rastogi

(2007) and extensively validated in Pahikkala et al. (2008) and Airolaa et al. (2011)—is a method for providing an unbiased estimate of the AUC, and the one we use in this article. In the leave-pair-out technique, every pairing between a negative example (i.e., a participant without MCI) and a positive example (i.e., a participant with MCI) is tested using a classifier trained on all of the remaining examples. The results of each positive/negative pair can be used to calculate the Wilcoxon-Mann-Whitney statistic as follows. Let $s(e)$ be the score of some example e; let P be the set of positive examples and N the set of negative examples; and let $[s(p) > s(n)]$ be 1 if true and 0 if false. Then:

$$AUC(s, P, n) = \frac{1}{|P||N|} \sum_{p \in P} \sum_{n \in N} [s(p) > s(n)] \tag{1}$$

Although this method is compute-intensive, it does provide an unbiased estimate of the AUC, whereas other cross-validation setups lead to biased estimates. Another benefit of using the AUC is that standard deviation can be calculated. The standard deviation for the AUC is calculated as follows, where AUC is abbreviated as A to improve readability:

$$\sigma_a^2 = \frac{A(1-A) + (|P|-1)(\frac{A}{2-A} - A^2) + (|N|-1)(\frac{2A^2}{1+A} - A^2)}{|P||N|} \tag{2}$$

4.2 Baseline Features

Previous work has shown that the WLM summary scores assigned during standard administration of the WLM, particularly in combination with other tests of verbal fluency and memory, are sensitive to the presence of MCI and other dementias (Storandt and Hill 1989; Petersen et al. 1999; Schmitt et al. 2000; Wang and Zhou 2002; Nordlund et al. 2005; Bennett et al. 2006; Price et al. 2009). We note, however, that the WLM test alone is not typically used as a diagnostic test. One of the goals of this work is to explore the utility of the standard WLM summary scores for diagnostic classification. A more ambitious goal is to demonstrate that using smaller units of information derived from story elements, rather than gross summary-level scores, can greatly improve diagnostic accuracy. Finally, we will show that using element-level scores automatically extracted from word alignments can achieve diagnostic classification accuracy comparable to that achieved using manually assigned scores. We therefore will compare the accuracy, measured in terms of AUC, of SVM classifiers trained on both summary-level and element-level WLM scores extracted from word alignments to the accuracy of classifiers built using a variety of alternative feature sets, both manually and automatically derived, shown in Table 3.

First, we consider the accuracy of classifiers using the expert-assigned WLM scores as features. For each of the 235 experimental participants, we generate two summary scores: one for the immediate retelling and one for the delayed retelling. The summary score ranges from 0, indicating that no elements were recalled, to 25, indicating that all elements were recalled. Previous work using manually assigned scores as features indicate that certain elements are more powerful in their ability to predict the presence of MCI (Prud'hommeaux 2012). In addition to the summary score, we therefore also provide the SVM with a vector of 50 story element-level scores: For each of the 25 elements in each of the two retellings per patient, there is a vector element with the value of 0 if the element was not recalled, or 1 if the element was recalled.

Table 3
Baseline classification accuracy results and standard deviation (s.d.).

Model	AUC (s.d.)
Manual WLM summary scores	73.3 (3.8)
Manual WLM element scores	81.3 (3.3)
LSA	74.8 (3.7)
Unigram overlap precision	73.3 (3.8)
BLEU	73.6 (3.8)
ROUGE-SU4	76.6 (3.6)
Exact match open-class summary score	74.3 (3.7)
Exact match open-class unigrams	76.4 (3.6)
MMSE	72.3 (3.8)

Classification accuracy with participants with MCI using these two manually derived feature sets is shown in Table 3.

We then present in Table 3 the classification accuracy of several summary-level features derived automatically from the WLM retellings, using standard NLP techniques for evaluating the similarity of two texts. We note that none of these features makes reference to the published WLM scoring guidelines or to the predefined element boundaries. Each of these feature sets contains two scores ranging between 0 and 1 for each participant, one for each of the two retellings: (1) cosine similarity between a retelling and the source narrative measured using LSA, proposed by Dunn et al. (2002) and calculated using the University of Colorado's online LSA interface (available at `http://lsa.colorado.edu/`) with the 300-factor ninth-grade reading level topic space; (2) unigram overlap precision of a retelling relative to the source, proposed by Hakkani-Tur, Vergyri, and Tur (2010); (3) BLEU, the *n*-gram overlap metric commonly used to evaluate the quality of machine translation output (Papineni et al. 2002); and (4) the F-measure for ROUGE-SU4, the *n*-gram overlap metric commonly used to evaluate automatic summarization output (Lin 2004). The remaining two automatically derived features are a set of binary scores corresponding to the exact match via `grep` of each of the open-class unigrams in the source narrative and a summary score thereof.

Finally, in order to compare the WLM with another standard psychometric test, we also show the accuracy of a classifier trained only on the expert-assigned manual scores for the MMSE (Folstein, Folstein, and McHugh 1975), a clinician-administered 30-point questionnaire that measures a patient's degree of cognitive impairment. Although it is widely used to screen for dementias such as Alzheimer's disease, the MMSE is reported not to be particularly sensitive to MCI (Morris et al. 2001; Ravaglia et al. 2005; Hoops et al. 2009). The MMSE is entirely independent of the WLM and, though brief (5–10 minutes), requires more time to administer than the WLM.

In Table 3, we see that the WLM-based features yield higher accuracy than the MMSE, which is notable given the role that the MMSE plays in dementia screening. In addition, although all of the automatically derived feature sets yield higher classification than the MMSE, the manually derived WLM element-level scores are by far the most accurate feature set for diagnostic classification. Summary-level statistics, whether derived manually using established scoring mechanisms or automatically using a variety of text-similarity metrics used in the NLP community, seem not to provide sufficient power to distinguish the two diagnostic groups. In the next several sections, we describe a method for accurately automatically extracting the identities of the recalled story

elements from WLM retellings via word alignment in order to try to achieve classification accuracy comparable to that of the manually assigned WLM story elements and higher than that of the other automatic scoring methods.

5. WLM Scoring Via Alignment

The approach presented here for automatic scoring of the WLM subtest relies on word alignments of the type used in machine translation for building phrased-based translation models. The motivation for using word alignment is the inherent similarity between narrative retelling and translation. In translation, a sentence in one language is converted into another language; the translation will have different words presented in a different order, but the meaning of the original sentence will be preserved. In narrative retelling, the source narrative is "translated" into the idiolect of the individual retelling the story. Again, the retelling will have different words, possibly presented in a different order, but at least some of the meaning will be preserved. We will show that although the algorithm for extracting scores from the alignments is simple, the process of getting high quality word alignments from the corpora of narrative retellings is challenging.

Although researchers in other NLP tasks that rely on alignments, such as textual entailment and summarization, sometimes eschew the sort of word-level alignments that are used in machine translation, we have no a priori reason to believe that this sort of alignment will be inadequate for the purposes of scoring narrative retellings. In addition, unlike many of the alignment algorithms proposed for tasks such as textual entailment, the methods for unsupervised word alignment used in machine translation require no external resources or hand-labeled data, making it simple to adapt our automated scoring techniques to new scenarios. We will show that the word alignment algorithms used in machine translation, when modified in particular ways, provide sufficient information for highly accurate scoring of narrative retellings and subsequent diagnostic classification of the individuals generating those retellings.

5.1 Example Alignment

Figure 4 shows a visual grid representation of a manually generated word alignment between the source narrative shown in Figure 1 on the vertical axis and the example WLM retelling in Figure 2 on the horizontal axis. Table 4 shows the word-index-to-word-index alignment, in which the first index of each sentence is 0 and in which null alignments are not shown.

When creating these manual alignments, the labelers assigned the "possible" denotation under one of these two conditions: (1) when the alignment was ambiguous, as outlined in Och and Ney (2003); and (2) when a particular word in the retelling was a logical alignment to a word in the source narrative, but it would not have been counted as a permissible substitution under the published scoring guidelines. For this reason, we see that *Taylor* and *sixty-seven* are considered to be possible alignments because although they are logical alignments, they are not permissible substitutions according to the published scoring guidelines. Note that the word *dollars* is considered to be only a possible alignment, as well, since the element *fifty-six dollars* is not correctly recalled in this retelling under the standard scoring guidelines. In Figure 4, sure alignments are marked in black and possible alignments are marked in gray. In Figure 5, sure alignments are marked with *S* and possible alignments are marked with *P*.

Manually generated alignments like this one are the gold standard against which any automatically generated alignments can be compared to determine the accuracy

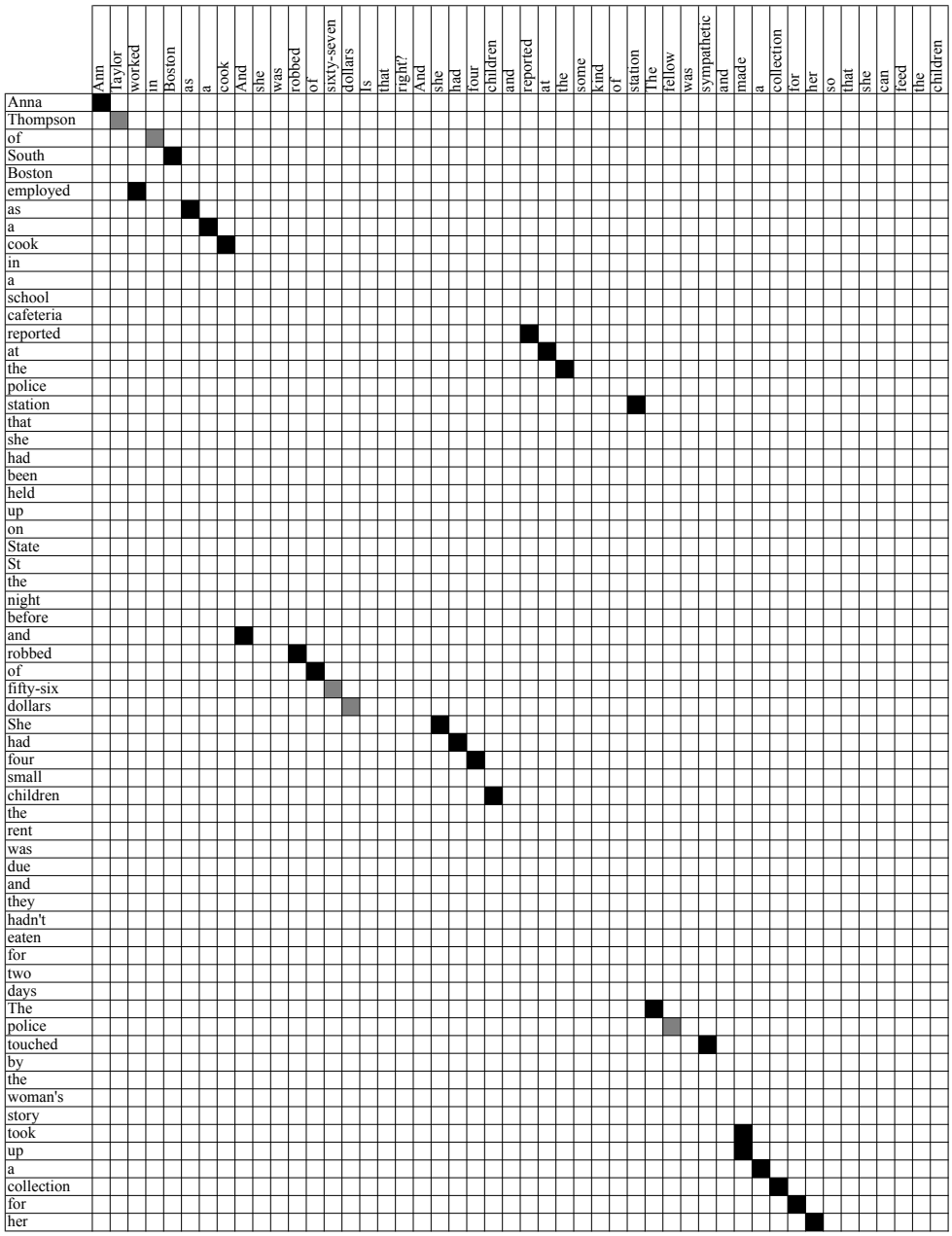

Figure 4
Visual representation of word alignment of source narrative and sample narrative.

of the alignment. From an accurate word-to-word alignment, the identities of the story elements used in a retellings can be accurately extracted, and from that set of story elements, the score that is assigned under the standard scoring procedure can be calculated.

Table 4
Sure (S) and Possible (P) index-to-index word alignment of the narrative in Figure 2.

Source	Retelling	S/P
anna(0)	ann(0)	S
thompson(1)	taylor(1)	P
employed(5)	worked(2)	S
of(2)	in(3)	P
boston(4)	boston(4)	S
as(6)	as(5)	S
a(7)	a(6)	S
cook(8)	cook(7)	S
robbed(31)	robbed(11)	S
of(32)	of(12)	S
fifty-six(33)	sixty-seven(13)	P
dollars(34)	dollars(14)	P
she(35)	she(19)	S
had(36)	had(20)	S
four(37)	four(21)	S
children(39)	children(22)	S
reported(13)	reported(24)	S
at(14)	at(25)	S
the(15)	the(26)	S
station(17)	station(30)	S
the(52)	the(31)	S
police(53)	fellow(32)	P
touched(54)	sympathetic(34)	S
took(59)	made(36)	S
up(60)	made(36)	S
a(61)	a(37)	S
collection(62)	collection(38)	S
for(63)	for(39)	S
her(64)	her(40)	S

5.2 Story Element Extraction and Scoring

As described earlier, the published scoring guidelines for the WLM specify the source words that compose each story element. Figure 5 displays the source narrative with the element IDs ($A - Y$) and word IDs ($1 - 65$) explicitly labeled. Element Q, for instance, consists of the words 39 and 40, *small children*.

Using this information, we can determine which story elements were used in a retelling from the alignments as follows: for each word in the source narrative, if that word is aligned to a word in the retelling, the story element that it is associated with is considered to be recalled. For instance, if there is an alignment between the retelling word *sympathetic* and the source word *touched*, the story element *touched by the woman's story* would be counted as correctly recalled. Note that in the WLM, every word in the source narrative is part of one of the story elements. Thus, when we convert alignments to scores in the way just described, any alignment can generate a story element. This is true even for an alignment between function words such as *the* and *of*, which would be unlikely individually to indicate that a story element had been recalled. To avoid such scoring errors, we disregard any word alignment pair containing a function word from

[*A* anna$_0$] [*B* thompson$_1$] [*C* of$_2$ south$_3$] [*D* boston$_4$] [*E* employed$_5$] [*F* as$_6$ a$_7$ cook$_8$] [*G* in$_9$ a$_{10}$ school$_{11}$] [*H* cafeteria$_{12}$] [*I* reported$_{13}$] [*J* at$_{14}$ the$_{15}$ police$_{16}$] [*K* station$_{17}$] [*L* that$_{18}$ she$_{19}$ had$_{20}$ been$_{21}$ held$_{22}$ up$_{23}$] [*M* on$_{24}$ state$_{25}$ street$_{26}$] [*N* the$_{27}$ night$_{28}$ before$_{29}$] [*O* and$_{30}$ robbed$_{31}$ of$_{32}$] [*P* fifty-six$_{33}$ dollars$_{34}$] [*Q* she$_{35}$ had$_{36}$ four$_{37}$] [*R* small$_{38}$ children$_{39}$] [*S* the$_{40}$ rent$_{41}$ was$_{42}$ due$_{43}$] [*T* and$_{44}$ they$_{45}$ had$_{46}$ n't$_{47}$ eaten$_{48}$] [*U* for$_{49}$ two$_{50}$ days$_{51}$] [*V* the$_{52}$ police$_{53}$] [*W* touched$_{54}$ by$_{55}$ the$_{56}$ woman's$_{57}$ story$_{58}$] [*X* took$_{59}$ up$_{60}$ a$_{61}$ collection$_{62}$] [*Y* for$_{63}$ her$_{64}$]

Figure 5
Text of WLM narrative with story element bracketing and word IDs.

the source narrative. The two exceptions to this rule are the final two words, *for her*, which are not content words but together make a single story element.

Recall that in the manually derived word alignments, certain alignment pairs were marked as *possible* if the word in the retelling was logically equivalent to the word in the source but was not a permissible substitute according to the published scoring guidelines. When extracting scores from a manual alignment, only *sure* alignments are considered. This enables us to extract scores from a manual word alignment with 100% accuracy. The *possible* manual alignments are used only for calculating alignment error rate (AER) of an automatic word alignment model.

From the list of story elements extracted in this way, the summary score reported under standard scoring guidelines can be determined simply by counting the number of story elements extracted. Table 5 shows the story elements extracted from the manual word alignment in Table 4.

5.3 Word Alignment Data

The WLM immediate and delayed retellings for all of the 235 experimental participants and the 48 retellings from participants in the larger study who were not eligible for the present study were transcribed at the word level. Partial words, punctuation, and pause-fillers were excluded from all transcriptions used for this study. The retellings were manually scored according to published guidelines. In addition, we manually

Table 5
Alignment from Table 4, excluding function words, with associated story element IDs.

Element ID	Source word : Retelling word
A	anna(0) : ann(0)
E	employed(5) : worked(2)
D	boston(4) : boston(4)
F	cook(8) : cook(7)
O	robbed(31) : robbed(11)
Q	four(37) : four(21)
R	children(39) : children(22)
I	reported(13) : reported(24)
K	station(17) : station(30)
W	touched(54) : sympathetic(34)
X	took(59) : made(36)
X	collection(62) : collection(38)
Y	for(63) : for(39)
Y	her(64) : her(40)

produced word-level alignments between each retelling and the source narrative presented. These manual alignments were used to evaluate the word alignment quality and never to train the word alignment model.

Word alignment for phrase-based machine translation typically takes as input a sentence-aligned parallel corpus or bi-text, in which a sentence on one side of the corpus is a translation of the sentence in that same position on the other side of the corpus. Because we are interested in learning how to align words in the source narrative to words in the retellings, our primary parallel corpus must consist of source narrative text on one side and retelling text on the other. Because the retellings contain omissions, reorderings, and embellishments, we are obliged to consider the full text of the source narrative and of each retelling to be a "sentence" in the parallel corpus.

We compiled three parallel corpora to be used for the word alignment experiments:

- **Corpus 1:** A 518-line source-to-retelling corpus consisting of the source narrative paired with each of the two retellings from the 235 experimental participants as well as the 48 retellings from ineligible individuals.
- **Corpus 2:** A 268,324-line pairwise retelling-to-retelling corpus, consisting of every possible pairwise combination of the 518 available retellings.
- **Corpus 3:** A 976-line word identity corpus, consisting of every word that appears in any retelling and the source narrative paired with itself.

The explicit parallel alignments of word identities that compose Corpus 3 are included in order to encourage the alignment of a word in a retelling to that same word in the source, if it exists.

The word alignment techniques that we use are unsupervised. Other than the transcriptions themselves, no manually generated data is used to build the word alignment models. Therefore, as in the case with most experiments involving word alignment, we build a model for the data we wish to evaluate using that same data. We do, however, use the 48 retellings from the individuals who were not experimental participants as a development set for tuning the various parameters of our word alignment system, which are described in the following.

5.4 Baseline Alignment

We begin by building two word alignment models using the Berkeley aligner (Liang, Taskar, and Klein 2006), a state-of-the-art word alignment package that relies on IBM Models 1 and 2 (Brown et al. 1993) and an HMM. We chose to use the Berkeley aligner, rather than the more widely used Giza++ alignment package, for this task because its joint training and posterior decoding algorithms yield lower alignment error rates on most data sets (including the data set used here [Prud'hommeaux and Roark 2011]) and because it offers functionality for testing an existing model on new data and, more crucially, for outputting posterior probabilities. The smaller of our two Berkeley-generated models is trained on Corpus 1 (the source-to-retelling parallel corpus described earlier) and ten copies of Corpus 3 (the word identity corpus). The larger model is trained on Corpus 1, Corpus 2 (the pairwise retelling corpus), and 100 copies of Corpus 3. Both models are then tested on the 470 retellings from our 235 experimental participants. In addition, we use both models to align every retelling to every other retelling so that we will have all pairwise alignments available for use in the graph-based model presented in the next section.

We note that the Berkeley aligner occasionally fails to return an alignment for a sentence pair, either because one of the sentences is too long or because the time required to perform the necessary calculations exceeds some maximum allotted time. In these cases, in order to generate alignments for all retellings and to build a complete graph that includes all retellings, we back off to the alignments and posteriors generated by IBM Model 1.

The first two rows of Table 6 show the precision, recall, and alignment error rate (AER) (Och and Ney 2003) for these two Berkeley aligner models. We note that although the AER for the larger model is lower, the time required to train the model is significantly longer. The alignments generated by the Berkeley aligner serve not only as a baseline for comparison of word alignment quality but also as a springboard for the novel graph-based method of alignment we will now discuss.

5.5 Graph-Based Refinement

Graph-based methods, in which paths or **random walks** are traced through an interconnected graph of nodes in order to learn more about the nodes themselves, have been used for NLP tasks in information extraction and retrieval, including Web-page ranking (PageRank; Page et al. 1999) and extractive summarization (LexRank; Erkan and Radev 2004; Otterbacher, Erkan, and Radev 2009). In the PageRank algorithm, the nodes of the graph are Web pages and the edges connecting the nodes are the hyperlinks leading from those pages to other pages. The nodes in the LexRank algorithm are sentences in a document and the edges are the similarity scores between those sentences. The number of times that a particular node is visited in a random walk reveals information about the importance of that node and its relationship to the other nodes. In many applications of random walks, the goal is to determine which node is the most central or has the highest prestige. In word alignment, however, the goal is to learn new relationships and strengthen existing relationships between words in a retelling and words in the source narrative.

In the case of our graph-based method for word alignment, each node represents a word in one of the retellings or in the source narrative. The edges are the normalized posterior-weighted alignments that the Berkeley aligner proposes between each word and (1) words in the source narrative, and (2) words in the other retellings. We generate these edges by using an existing baseline alignment model to align every retelling to every other retelling and to the source narrative. The posterior probabilities produced by the baseline alignment model serve as the weights on the edges. At each step in the walk, the choice of the next destination node can be determined according to

Table 6
Word alignment performance.

Model	P	R	AER
Berkeley-Small	72.3	78.5	24.8
Berkeley-Large	79.0	79.4	20.9
Graph-based-Small	77.9	81.2	20.6
Graph-based-Large	85.4	76.9	18.9

UPDATESRCDIST(R, S, $\mathcal{P}$, $\mathcal{Q}$, λ, N)
Inputs: retelling words R; source words S; transition multinomial model $\mathcal{P}$: R $\rightarrow$ R; transition multinomial model $\mathcal{Q}$: R $\rightarrow$ S; parameters λ, N
Output: updated transition multinomial model matrix $\mathcal{Q}$: R $\rightarrow$ S

$\mathcal{Q}' \leftarrow$ zeros($|R|$, $|S|$) ▷ Initialize $|R| \times |S|$ matrix with zeros.
for r in R **do**
  for $i = 1$ to N **do**
    $r' \leftarrow r$
    while RAND[0, 1] $\leq \lambda$ **do** ▷ until random probability $> \lambda$
      $r' \leftarrow$ SAMPLE($\mathcal{P}(r', :)$) ▷ sample destination retelling word from r'
    $s \leftarrow$ SAMPLE($\mathcal{Q}(r', :)$) ▷ sample destination source word from r'
    $\mathcal{Q}'[r, s] \leftarrow \mathcal{Q}'[r, s] + \frac{1}{N}$ ▷ accumulate evidence in new matrix
return $\mathcal{Q}'$

Figure 6
Pseudocode for updating the multinomial model for transitioning from retelling words to source words.

the strength of the outgoing edges, as measured by the posterior probability of that alignment.

Starting at a word in one of the retellings, represented by a node in the graph, the algorithm can walk from that node either to another retelling word in the graph to which it is aligned or to a word in the source narrative to which it is aligned. At each step in the walk, there is an empirically derived probability, λ, that sets the likelihood of transitioning to another retelling word versus a word in the source narrative. This probability functions similarly to the damping factor used in PageRank and LexRank, although its purpose is quite different. Once the decision whether to walk to a retelling word or source word has been made, the destination word itself is chosen according to the weights, which are the posterior probabilities assigned by the baseline alignment model. When the walk arrives at a source narrative word, that particular random walk ends, and the count for that source word as a possible alignment for the input retelling word is incremented by one.

For each word in each retelling, we perform 1,000 of these random walks, thereby generating a distribution for each retelling word over all of the words in the source narrative. The new alignment for the word is the source word with the highest frequency in that distribution. Pseudocode for this algorithm is provided in Figure 6.

Consider the following excerpts of five of the retellings. In each excerpt, the word that should align to the source word *touched* is rendered in bold:

the police were so **moved** by the story that they took up a collection for her

the fellow was **sympathetic** and made a collection for her so that she can feed the children

the police were **touched** by their story so they took up a collection

the police were so **impressed** with her story they took up a collection

the police felt **sorry** for her and took up a collection

Figure 7 presents a small idealized subgraph of the pairwise alignments of these five retellings. The arrows represent the alignments proposed by the Berkeley aligner between the relevant words in the retellings and their alignment (or lack of alignment) to the word *touched* in the source narrative. Thin arrows indicate alignment edges in the graph between retelling words, and bold arrows indicate alignment edges between retelling words and words in the source narrative. Words in the retellings are rendered as nodes with a single outline, and words in the source are rendered as nodes with a double outline.

We see that a number of these words were not aligned to the correct source word, *touched*. They are all, however, aligned to other retelling words that are in turn eventually aligned to the source word. Starting at any of the nodes in the graph, it is possible to walk from node to node and eventually reach the correct source word. Although *sympathetic* was not aligned to *touched* by the Berkeley aligner, its correct alignment can be recovered from the graph by following the path through other retelling words. After hundreds or thousands of random walks on the graph, evidence for the correct alignment will accumulate.

The approach as described might seem most beneficial to a system in need of improvements to recall rather than precision. Our baseline systems, however, are already favoring recall over precision. For this reason we include the NULL word in the list of words in the source narrative. We note that most implementations of both IBM Model 1 and HMM-based alignment also model the probability of aligning to a hidden word, NULL. In word alignment for machine translation, alignment to NULL usually indicates that a word in one language has no equivalent in the other language because the two languages express the same idea or construction in a slightly different way. Romance languages, for instance, often use prepositions before infinitival complements (e.g., Italian *cerco di ridere*) when English does not (e.g., *I try to laugh*). In the alignment of narrative retellings, however, alignment to NULL often indicates that the word in question is part of an aside or a piece of information that was not expressed in the source narrative.

Any retelling word that is not aligned to a source word by the baseline alignment system will implicitly be aligned to the hidden source word NULL, guaranteeing that every retelling word has at least one outgoing alignment edge and allowing us to

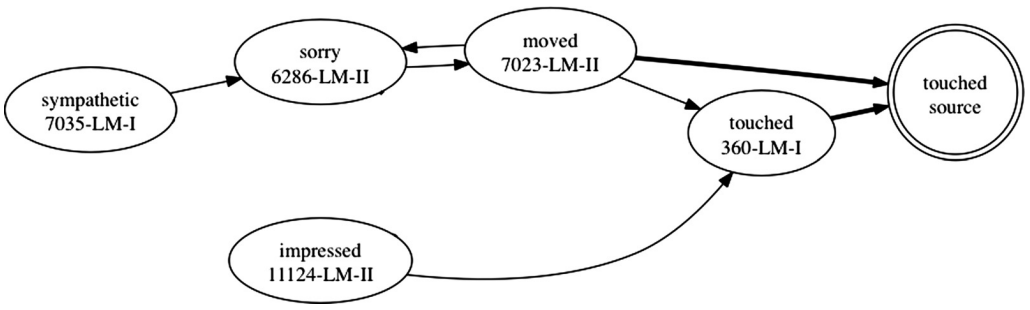

Figure 7
Subgraph of the full pairwise and source-to-retelling alignment. Thin arrows represent alignments from retelling words to other retelling words. Bold arrows indicate alignments to words from the source narrative, which are rendered as nodes with a double outline.

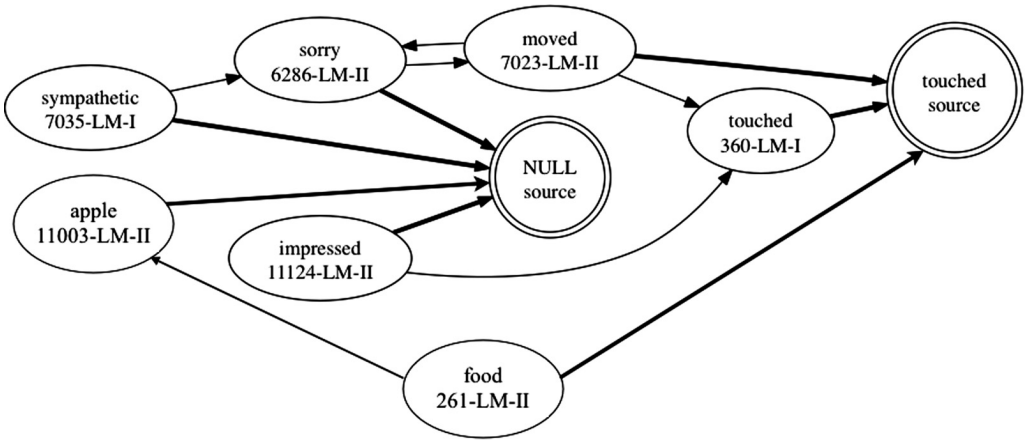

Figure 8
Subgraph of the full pairwise and source-to-retelling alignment including the NULL word. Thin arrows represent alignments from retelling words to other retelling words. Bold arrows indicate alignments to words from the source narrative, which are rendered as nodes with a double outline.

model the likelihood of being unaligned. A word that was unaligned by the original system can remain unaligned. A word that should have been left unaligned but was mistakenly aligned to a source word by the original system can recover its correct (lack of) alignment by following an edge to another retelling word that was correctly left unaligned (i.e., aligned to NULL). Figure 8 shows the graph in Figure 7 with the addition of the NULL node and the corresponding alignment edges to that node. This figure also includes two new retelling words, *food* and *apple*, and their respective alignment edges. Here we see that although the retelling word *food* was incorrectly aligned to the source word *touched* by the baseline system, its correct alignment to NULL can be recovered by traversing the edge to retelling word *apple* and from there, the edge to the source word NULL.

The optimal values for the following two parameters for the random walk must be determined: (1) the value of λ, the probability of walking to a retelling word node rather than a source word, and (2) the posterior probability threshold for including a particular edge in the graph. We optimize these parameters by testing the output of the graph-based approach on the development set of 48 retellings from the individuals who were not eligible for the study, discussed in Section 5.3. Recall that these additional retellings were included in the training data for the alignment model but were not included in the test set used to evaluate its performance. Tuning on this set of retellings therefore introduces no additional words, out-of-vocabulary words, or other information to the graph, while preventing overfitting.

The posterior threshold is set to 0.5 in the Berkeley aligner's default configuration, and we found that this value did indeed yield the lowest AER for the Berkeley aligner on our data. When building the graph using Berkeley alignments and posteriors, however, we can adjust the value of this threshold to optimize the AER of the alignments produced via random walks. Using the development set of 48 retellings, we determined that the AER is minimized when the value of λ is 0.8 and the alignment inclusion posterior threshold is 0.5.

5.6 Word Alignment Evaluation

Recall the two baseline alignment models generated by the Berkeley aligner, described in Section 5.4: (1) the small Berkeley model, trained on Corpus 1 (the source-to-retelling corpus) and 10 instances of Corpus 3 (the word identity corpus), and (2) the large Berkeley model (trained on Corpus 1, Corpus 2, the full pairwise retelling-to-retelling corpus, and 100 instances of Corpus 3). Using these models, we generate full retelling-to-retelling alignments, on which we can then build two graph-based alignment models: the small graph-based model and the large graph-based model.

The manual gold alignments for the 235 experimental participants were evaluated against the alignments produced by each of the four models. Table 6 presents the precision, recall, and AER for the alignments of the experimental participants. Not surprisingly, the larger models yield lower error rates than the smaller models. More interestingly, each graph-based model outperforms the Berkeley model of the corresponding size by a large margin. The performance of the small graph-based model is particularly remarkable because it yields an AER superior to the large Berkeley model while requiring significantly fewer computing resources. Each of the graph-based models generated the full set of alignments in only a few minutes, whereas the large Berkeley model required 14 hours of training.

6. Scoring Evaluation

The element-level scores induced, as described in Section 5.2, from the four word alignments for all 235 experimental participants were evaluated against the manual per-element scores. We report the precision, recall, and F-measure for all four alignment models in Table 7. In addition, we report Cohen's kappa as a measure of reliability between our automated scores and the manually assigned scores. We see that as AER improves, scoring accuracy also improves, with the large graph-based model outperforming all other models in terms of precision, F-measure, and inter-rater reliability. The scoring accuracy levels reported here are comparable to the levels of inter-rater agreement typically reported for the WLM, and reliability between our automated scores and the manual scores, as measured by Cohen's kappa, is well within the ranges reported in the literature (Johnson, Storandt, and Balota 2003). As will be shown in the following section, scoring accuracy is important for achieving high classification accuracy of MCI.

Table 7
Scoring accuracy results.

Model	P	R	F	κ
Berkeley-Small	87.2	88.9	88.0	76.1
Berkeley-Large	86.8	90.7	88.7	77.1
Graph-Small	84.7	93.6	88.9	76.9
Graph-Big	88.8	89.3	89.1	78.3

7. Diagnostic Classification

As discussed in Section 2, poor performance on the WLM test is associated with MCI. We now use the scores we have extracted from the word alignments as features with an SVM to perform diagnostic classification for distinguishing participants with MCI from those without, as described in Section 4.1.

Table 8 shows the classification results for the scores derived from the four alignment models along with the classification results using the examiner-assigned manual scores, the MMSE, and the four alternative automated scoring approaches described in Section 4.2. It appears that, in all cases, the per-element scores are more effective than the summary scores in classifying the two diagnostic groups. In addition, we see that our automated scores have classificatory power comparable to that of the manual gold scores, and that as scoring accuracy increases from the small Berkeley model to the graph-based models and bigger models, classification accuracy improves. This suggests both that accurate scores are crucial for accurate classification and that pursuing even further improvements in word alignment is likely to result in improved diagnostic differentiation. We note that although the large Berkeley model achieved the highest classification accuracy of the automated methods, this very slight margin of difference may not justify its significantly greater computational requirements.

In addition to using summary scores and element-level scores as features for the story-element based models, we also perform feature selection over both sets of features using the chi-square statistic. Feature selection is performed separately on each training set for each fold in the cross-validation to avoid introducing bias from the testing example. We train and test the SVM using 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. We note that over all of the folds, only 8 of the 50 features ever appeared among the seven most informative.

In all cases, the per-element scores are more effective than the summary scores in classifying the two diagnostic groups, and performing feature selection results in improved classification accuracy. All of the element-level feature sets automatically extracted from alignments outperform the MMSE and all of the alternative automatic

Table 8
Classification accuracy results measured by AUC (standard deviation).

Model	Summary scores	Element scores	Element subset
Berkeley-Small	73.7 (3.74)	77.9 (3.52)	80.3 (3.4)
Berkeley-Big	75.1 (3.67)	79.2 (3.45)	81.2 (3.3)
Graph-Small	74.2 (3.71)	78.9 (3.47)	80.0 (3.4)
Graph-Big	74.8 (3.69)	78.6 (3.49)	81.6 (3.3)
Manual Scores	73.3 (3.76)	81.3 (3.32)	82.1 (3.3)
MMSE	72.3 (3.8)	n/a	n/a
LSA	74.8 (3.7)	n/a	n/a
BLEU	73.6 (3.8)	n/a	n/a
ROUGE-SU4	76.6 (3.6)	n/a	n/a
Unigram overlap precision	73.3 (3.8)	n/a	n/a
Exact match open-class	74.3 (3.7)	76.4 (3.6)	n/a

scoring procedures, which suggests that the extra complexity required to extract element-level features is well worth the time and effort.

We note that the final classification results for all four alignment models are not drastically different from one another, despite the large reductions in word alignment error rate and improvements in scoring accuracy observed in the larger models and graph-based models. This seeming disconnect between word alignment accuracy and downstream application performance has also been observed in the machine translation literature, where reductions in AER do not necessarily lead to meaningful increases in BLEU, the widely accepted measure of machine translation quality (Ayan and Dorr 2006; Lopez and Resnik 2006; Fraser and Marcu 2007). Our results, however, show that a feature set consisting of manually assigned WLM scores yields the highest classification accuracy of any of the feature sets evaluated here. As discussed in Section 5.2, our WLM score extraction method is designed such that element-level scores can be extracted with perfect accuracy from a perfect word alignment. Thus, the goal of seeking perfect or near-perfect word alignment accuracy is worthwhile because it will necessarily result in perfect or near-perfect scoring accuracy, which in turn is likely to yield classification accuracy approaching that of manually assigned scores.

8. Application to Task with Non-linguistic Reference

As we discussed earlier, one of the advantages of using an unsupervised method of scoring is the resulting generalizability to new data sets, particularly those generated from a non-linguistic stimulus. The Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan 1972), an instrument widely used to diagnose aphasia in adults, includes one such task, popularly known as the cookie theft picture description task. In this test, the person views a drawing of a lively scene in a family's kitchen and must tell the examiner about all of the actions they see in the picture. The picture is reproduced below in Figure 9.

Describing visually presented material is quite different from a task such as the WLM, in which language comprehension and memory play a crucial role. Nevertheless, the processing and language production demands of a picture description task may lead to differences in performance in groups with certain cognitive and language problems. In fact, it is widely reported that the picture descriptions of seniors with dementia of the Alzheimer's type differ from those of typically aging seniors in terms of information content (Hier, Hagenlocker, and Shindler 1985; Gilesa, Patterson, and Hodge 1996). Interestingly, this reduction in information is not necessarily accompanied by a reduction in the amount of language produced. Rather, it seems that seniors with Alzheimer's dementia tend to include redundant information, repetitions, intrusions, and revisions that result in language samples of length comparable to that of typically aging seniors.

TalkBank (MacWhinney 2007), the online database of audio and transcribed speech, has made available the DementiaBank corpus of descriptions of the cookie theft picture by hundreds of individuals, some of whom have one of a number of types of dementia, including MCI, vascular dementia, possible Alzheimer's disease, and probable Alzheimer's disease. From this corpus were selected a subset of individuals without dementia and a subset with probable Alzheimer's disease. We limit the set of descriptions to those with more than 25 but fewer than 100 words, yielding 130 descriptions for each diagnostic group. There was no significant difference in description word count between the two diagnostic groups.

The first task was to generate a source description to which all other narratives should be aligned. Working under the assumption that the control participants would

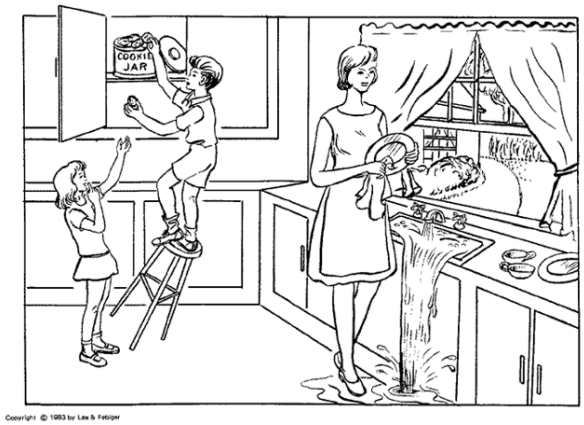

Figure 9
BDAE cookie theft picture.

produce good descriptions, we calculated the BLEU score of every pair of descriptions from the control group. The description with the highest average pairwise BLEU score was selected as the source description. After confirming that this description did in fact contain all of the action portrayed in the picture, we removed all extraneous conversational asides from the description in order to ensure that it contained all and only information about the picture. The selected source description is as follows:

> The boy is getting cookies out of the cookie jar. And the stool is just about to fall over. The little girl is reaching up for a cookie. And the mother is drying dishes. The water is running into the sink and the sink is running over onto the floor. And that little girl is laughing.

We then built an alignment model on the full pairwise description parallel corpus (260^2 = 67,600 sentences) and a word identity corpus consisting of each word in each description reproduced 100 times. Using this trained model, which corresponds to the large Berkeley model that achieved the highest classification accuracy for the WLM data, we then aligned every description to the artificial source description. We also built a graph-based alignment model using these alignments and the parameter settings that maximized word alignment accuracy in the WLM data. Because the artificial source description is not a true linguistic reference for this task, we did not produce manual word alignments against which the alignment quality could be evaluated and against which the parameters could be tuned. Instead, we evaluated only the downstream application of diagnostic classification.

The method for scoring the WLM relies directly on the predetermined list of story elements, whereas the cookie theft picture description administration instructions do not include an explicit set of items that must be described. Recall that the automated scoring method we propose uses only the open-class or content words in the source narrative. In order to generate scores for the descriptions, we propose a scoring technique that considers each content word in the source description to be its own story element. Any word in a retelling that aligns to one of the content words in the source narrative is considered to be a match for that content word element. This results in a large number of elements, but it allows the scoring method to be easily adapted to other

Table 9
Classification accuracy for probable Alzheimer's disease using BDAE picture description features, and standard deviation (s.d.).

Feature Set	AUC (s.d.)
Unigram precision	63.0 (3.4)
BLEU	70.1 (3.2)
Content word summary score (graph-based)	70.4 (3.2)
Content word word-level scores (graph-based)	82.3 (2.6)
Content word summary score (Berkeley)	67.6 (3.3)
Content word word-level scores (Berkeley)	83.2 (2.5)

narrative production scenarios that similarly do not have explicit scoring guidelines. Using these scores as features, we again used an SVM to classify the two diagnostic groups, typically aging and probable Alzheimer's disease, and evaluated the classifier using leave-pair-out validation.

Table 9 shows the classification results using the content word scoring features produced using the Berkeley aligner alignments and the graph-based alignments. These can be compared to classification results using the summary similarity metrics BLEU and unigram precision. We see that using word-level features, regardless of which alignment model they are extracted from, results in significantly higher classification accuracy than both the simple similarity metrics and the summary scores. The alignment-based scoring approach yields features with remarkably high classification accuracy given the somewhat ad hoc selection of the source narrative from the set of control retellings.

These results demonstrate the flexibility and utility of the alignment-based approach to scoring narratives. Not only can it be adapted to other narrative retelling instruments, but it can relatively trivially be adapted to instruments that use non-linguistic stimuli for elicitation. All that is needed to build an alignment model is a sufficiently large collection of retellings of the same narrative or descriptions of the same picture. Procuring such a collection of descriptions or retellings can be done easily outside a clinical setting using a platform such as Amazon's Mechanical Turk. No hand-labeled data, outside lexical resources, prior knowledge of the content of the story, or existing scoring guidelines are required.

9. Conclusions and Future Work

The work presented here demonstrates the utility of adapting NLP algorithms to clinically elicited data for diagnostic purposes. In particular, the approach we describe for automatically analyzing clinically elicited language data shows promise as part of a pipeline for a screening tool for mild cognitive impairment. The methods offer the additional benefit of being general and flexible enough to be adapted to new data sets, even those without existing evaluation guidelines. In addition, the novel graph-based approach to word alignment results in large reductions in alignment error rate. These reductions in error rate in turn lead to human-level scoring accuracy and improved diagnostic classification.

The demand for simple, objective, and unobtrusive screening tools for MCI and other neurodegenerative and neurodevelopmental disorders will continue to grow

as the prevalence of these disorders increases. Although high-level measures of text similarity used in other NLP applications, such as machine translation, do achieve reasonable classification accuracy when applied to the WLM narrative data, the work presented here indicates that automated methods that approximate manual element-level scoring procedures yield superior results.

Although the results are quite robust, several enhancements and improvements can be made. First, although we were able to achieve decent word alignment accuracy, especially with our graph-based approach, many alignment errors remain. Exploration of the graph used here reveals that many correct alignments remain undiscovered, with an oracle AER of 11%. One clear weakness is the selection of only a single alignment from the distribution of source words at the end of 1,000 walks, since this does not allow for one-to-many mappings. We would also like to experiment with including non-directional edges and outgoing edges on source words.

In our future work, we also plan to examine longitudinal data for individual participants to see whether our techniques can detect subtle differences in recall and coherence between a recent retelling and a series of earlier baseline retellings. Because the CDR, the dementia staging system often used to identify MCI, relies on observed changes in cognitive function over time, longitudinal analysis of performance on narrative retelling and picture description tasks might be the most promising application for this approach to analyzing clinically elicited language data.

Acknowledgments

This research was conducted while both authors were at the Center for Spoken Language Understanding at the Oregon Health and Science University, in Portland, Oregon. This work was supported in part by NSF grant BCS-0826654 and NIH NIDCD grants R01DC012033-01 and R01DC007129. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the NIH or NSF. Some of the results reported here appeared previously in Prud'hommeaux and Roark (2012) and the first author's dissertation (Prud'hommeaux 2012). We thank Jan van Santen, Richard Sproat, and Chris Callison-Burch for their valuable input and the clinicians at the OHSU Layton Center for their care in collecting the data.

References

Airolaa, Antti, Tapio Pahikkalaa, Willem Waegemanc, Bernard De Baetsc, and Tapio Salakoskia. 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, 55(4):1828–1844.

Artero, Sylvain, Mary Tierney, Jacques Touchon, and Karen Ritchie. 2003. Prediction of transition from cognitive impairment to senile dementia: A prospective, longitudinal study. *Acta Psychiatrica Scandinavica*, 107(5):390–393.

Ayan, Necip Fazil and Bonnie J. Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Sydney.

Bennett, D. A., J. A. Schneider, Z. Arvanitakis, J. F. Kelly, N. T. Aggarwal, R. C. Shah, and R. S. Wilson. 2006. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology*, 66(12):1837–1844.

Bishop, Dorothy and Chris Donlan. 2005. The role of syntax in encoding and recall of pictorial narratives: Evidence from specific language impairment. *British Journal of Developmental Psychology*, 23(1):25–46.

Brown, Peter, Vincent Della Pietra, Steven Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Busse, Anja, Anke Hensel, Uta Gühne, Matthias Angermeyer, and Steffi Riedel-Heller. 2006. Mild cognitive

impairment: Long-term course of four clinical subtypes. *Neurology*, 67(12):2176–2185.

Chang, Chih-Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27.

Chapman, Sandra, Hanna Ulatowska, Kristin King, Julene Johnson, and Donald McIntire. 1995. Discourse in early Alzheimer's disease versus normal advanced aging. *American Journal of Speech-Language Pathology*, 4:125–129.

Chenery, Helen J. and Bruce E. Murdoch. 1994. The production of narrative discourse in response to animations in persons with dementia of the Alzheimer's type: Preliminary findings. *Aphasiology*, 8(2):159–171.

Cortes, Corinna, Mehryar Mohri, and Ashish Rastogi. 2007. An alternative ranking problem for search engines. In *Proceedings of the 6th Workshop on Experimental Algorithms*, volume 4525 of *Lecture Notes in Computer Science*, pages 1–21.

Creamer, Scott and Maureen Schmitter-Edgecombe. 2010. Narrative comprehension in Alzheimer's disease: Assessing inferences and memory operations with a think-aloud procedure. *Neuropsychology*, 24(3):279–290.

de la Rosa, Gabriela Ramirez, Thamar Solorio, Manuel Montes y Gomez, Aquiles Iglesias, Yang Liu, Lisa Bedore, and Elizabeth Pena. 2013. Exploring word class n-grams to measure language development in children. In *Workshop on Biomedical Natural Language Processing (BIONLP 2013)*, pages 89–97, Sofia.

Diehl, Joshua J., Loisa Bennetto, and Edna Carter Young. 2006. Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, 34(1):87–102.

Dunn, John C., 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.

Ehrlich, Jonathan S., Loraine K. Obler, and Lynne Clark. 1997. Ideational and semantic contributions to narrative production in adults with dementia of the Alzheimer's type. *Journal of Communication Disorders*, 30(2):79–99.

Erkan, Günes and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Fan, Jerome, Suneel Upadhye, and Andrew Worster. 2006. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, 8:19–20.

Faraggi, David and Benjamin Reiser. 2002. Estimation of the area under the ROC curve. *Statistics in Medicine*, 21:3093–3106.

Folstein, M., 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.

Fraser, Alexander and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

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

Gabani, Keyur, Melissa Sherman, Thamar Solorio, and Yang Liu. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 46–55, Boulder, CO.

Galvin, James, Anne Fagan, David Holtzman, Mark Mintun, and John Morris. 2010. Relationship of dementia screening tests with biomarkers of Alzheimer's disease. *Brain*, 133:3290–3300.

Gilesa, Elaine, Karalyn Patterson, and John R. Hodge. 1996. Performance on the Boston cookie theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408.

Goodglass, H. and E. Kaplan. 1972. *Boston Diagnostic Aphasia Examination*. Lea and Febiger, Philadelphia, PA.

Hakkani-Tur, Dilek, Dimitra Vergyri, and Gokhan Tur. 2010. Speech-based automated cognitive status assessment. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*, pages 258–261, Makuhari.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

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

Hier, D., K. Hagenlocker, and A. Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, 25:117–133.

Hoops, S., S. Nazem, A. D. Siderowf, J. E. Duda, S. X. Xie, M. B. Stern, and D. Weintraub. 2009. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology*, 73(21):1738–1745.

Johnson, David K., Martha Storandt, and David A. Balota. 2003. Discourse analysis of logical memory recall in normal aging and in dementia of the Alzheimer type. *Neuropsychology*, 17(1):82–92.

Lautenschlager, Nicola T., John C. Dunn, Kathryn Bonney, Leon Flicker, and Osvaldo P. Almeida. 2006. Latent semantic analysis: An improved method to measure cognitive performance in subjects of non-English-speaking background. *Journal of Clinical and Experimental Neuropsychology*, 28:1381–1387.

Lehr, Maider, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1039–1042, Portland, OR.

Lehr, Maider, 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 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 211–220, Atlanta, GA.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 104–111, New York, NY.

Lin, Chin-Yiu. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona.

Lopez, Adam and Philip Resnik. 2006. Word-based alignment, phrase-based translation: What's the link. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 90–99, Cambridge.

Lysaker, Paul, Amanda Wickett, Neil Wilke, and John Lysaker. 2003. Narrative incoherence in schizophrenia: The absent agent-protagonist and the collapse of internal dialogue. *American Journal of Psychotherapy*, 57:153–166.

MacWhinney, Brian. 2007. The TalkBank Project. In J. C. Beal, K. P. Corrigan, and H. L. Moisl, editors, *Creating and Digitizing Language Corpora: Synchronic Databases, Vol.1*, pages 163–180, Palgrave-Macmillan, Houndmills.

Manly, Jennifer J., Ming Tang, Nicole Schupf, Yaakov Stern, Jean-Paul G. Vonsattel, and Richard Mayeux. 2008. Frequency and course of mild cognitive impairment in a multiethnic community. *Annals of Neurology*, 63(4):494–506.

Mitchell, Margaret. 1987. Scoring discrepancies on two subtests of the Wechsler memory scale. *Journal of Consulting and Clinical Psychology*, 55:914–915.

Morris, John. 1993. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, 43:2412–2414.

Morris, John, Martha Storandt, J. Phillip Miller, Daniel McKeel, Joseph Price, Eugene Rubin, and Leonard Berg. 2001. Mild cognitive impairment represents early-stage Alzheimer disease. *Archives of Neurology*, 58:397–405.

Norbury, Courtenay and Dorothy Bishop. 2003. Narrative skills of children with communication impairments. *International Journal of Language and Communication Disorders*, 38:287–313.

Nordlund, A., S. Rolstad, P. Hellstrom, M. Sjogren, S. Hansen, and A. Wallin. 2005. The Goteborg MCI study: Mild cognitive impairment is a heterogeneous condition. *Journal of Neurology, Neurosurgery and Psychiatry*, 76:1485–1490.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Otterbacher, Jahna, Günes Erkan, and Dragomir R. Radev. 2009. Biased LexRank: Passage retrieval using random walks with question-based priors. *Information Processing Management*, 45(1):42–54.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order

to the web. Technical Report 1999-66, Stanford InfoLab, Palo Alto, CA.

Pahikkala, Tapio, Antti Airola, Jorma Boberg, and Tapio Salakoski. 2008. Exact and efficient leave-pair-out cross-validation for ranking RLS. *The Second International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 1–8, Porvoo.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Petersen, Ronald, Glenn Smith, Stephen Waring, Robert Ivnik, Eric Tangalos, and Emre Kokmen. 1999. Mild cognitive impairment: Clinical characterizations and outcomes. *Archives of Neurology*, 56:303–308.

Petersen, Ronald C. 2011. Mild cognitive impairment. *The New England Journal of Medicine*, 364(23):2227–2234.

Plassman, Brenda L., Kenneth M. Langa, Gwenith G. Fisher, Steven G. Heeringa, David R. Weir, Beth Ofstedal, James R. Burke, Michael D. Hurd, Guy G. Potter, Willard L. Rodgers, David C. Steffens, John J. McArdle, Robert J. Willis, and Robert B. Wallace. 2008. Prevalence of cognitive impairment without dementia in the United States. *Annals of Internal Medicine*, 148:427–34.

Price, Joseph L., Daniel W. McKeel, Virginia D. Buckles, Catherine M. Roe, Chengjie Xiong, Michael Grundman, Lawrence A. Hansen, Ronald C. Petersen, Joseph E. Parisi, Dennis W. Dickson, Charles D. Smith, Daron G. Davis, Frederick A. Schmitt, William R. Markesbery, Jeffrey Kaye, Roger Kurlan, Christine Hulette, Brenda F. Kurland, Roger Higdon, Walter Kukull, and John C. Morris. 2009. Neuropathology of nondemented aging: Presumptive evidence for preclinical Alzheimer disease. *Neurobiology of Aging*, 30(7):1026–1036.

Prud'hommeaux, Emily and Brian Roark. 2011. Alignment of spoken narratives for automated neuropsychological assessment. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 484–489, Kona, HI.

Prud'hommeaux, Emily and Brian Roark. 2012. Graph-based alignment of narratives for automated neuropsychological assessment. In *Proceedings of the NAACL 2012 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 1–10, Montreal.

Prud'hommeaux, Emily and Masoud Rouhizadeh. 2012. Automatic detection of pragmatic deficits in children with autism. In *Proceedings of the 3rd Workshop on Child, Computer and Interaction*, pages 1–6, Portland, OR.

Prud'hommeaux, Emily Tucker. 2012. *Alignment of Narrative Retellings for Automated Neuropsychological Assessment*. Ph.D. thesis, Oregon Health and Science University.

Ravaglia, Giovanni, Paola Forti, Fabiola Maioli, Lucia Servadei, Mabel Martelli, Nicoletta Brunetti, Luciana Bastagli, and Erminia Mariani. 2005. Screening for mild cognitive impairment in elderly ambulatory patients with cognitive complaints. *Aging Clinical and Experimental Research*, 17(5):374–379.

Ridgway, James, Pierre Alquier, Nicolas Chopin, and Feng Liang. 2014. PAC-Bayesian AUC classification and scoring. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, 27:658–666.

Ritchie, Karen, Sylvaine Artero, and Jacques Touchon. 2001. Classification criteria for mild cognitive impairment: A population-based validation study. *Neurology*, 56:37–42.

Ritchie, Karen and Jacques Touchon. 2000. Mild cognitive impairment: Conceptual basis and current nosological status. *Lancet*, 355:225–228.

Roark, Brian, Margaret Mitchell, John-Paul Hosom, Kristina Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2081–2090.

Schmitt, F. A., D. G. Davis, D. R. Wekstein, C. D. Smith, J. W. Ashford, and W. R. Markesbery. 2000. Preclinical AD revisited: Neuropathology of cognitively normal older adults. *Neurology*, 55:370–376.

Shankle, William R., 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.
Solorio, Thamar and Yang Liu. 2008. Using language models to identify language impairment in Spanish-English bilingual children. In *Proceedings of the ACL 2008 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 116–117, Columbus, OH.
Storandt, Martha and Robert Hill. 1989. Very mild senile dementia of the Alzheimer's type: II. Psychometric test performance. *Archives of Neurology*, 46:383–386.
Tager-Flusberg, Helen. 1995. Once upon a ribbit: Stories narrated by autistic children. *British Journal of Developmental Psychology*, 13(1):45–59.
Tannock, Rosemary, Karen L. Purvis, and Russell J. Schachar. 1993. Narrative abilities in children with attention deficit hyperactivity disorder and normal peers. *Journal of Abnormal Child Psychology*, 21(1):103–17.
Tierney, Mary, Christie Yao, Alex Kiss, and Ian McDowell. 2005. Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. *Neurology*, 64:1853–1859.
Ulatowska, Hanna, Lee Allard, Adrienne Donnell, Jean Bristow, Sara M. Haynes, Adelaide Flower, and Alvin J. North. 1988. Discourse performance in subjects with dementia of the Alzheimer's type. In H.A. Whitaker, editor, *Neuropsychological Studies of Non-focal Brain Damage*. Springer-Verlag, New York.
United Nations. 2002. *World Population Ageing 1950–2050*. United Nations, New York.
Vuorinen, Elina, Matti Laine, and Juha Rinne. 2000. Common pattern of language impairment in vascular dementia and in Alzheimer disease. *Alzheimer Disease and Associated Disorders*, 14(2):81–86.
Wang, Qing-Song and Jiang-Ning Zhou. 2002. Retrieval and encoding of episodic memory in normal aging and patients with mild cognitive impairment. *Brain Research*, 924:113–115.
Wechsler, David. 1997. *Wechsler Memory Scale - Third Edition*. The Psychological Corporation, San Antonio, TX.
Zweig, Mark H. and Gregory Campbell. 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–577.