Graph-based alignment of narratives for automated neurological assessment

Emily T. Prud'hommeaux and Brian Roark
Center for Spoken Language Understanding
Oregon Health & Science University
{emilypx,roarkbr}@gmail.com

Abstract

Narrative recall tasks are widely used in neuropsychological evaluation protocols in order to detect symptoms of disorders such as autism, language impairment, and dementia. In this paper, we propose a graph-based method commonly used in information retrieval to improve word-level alignments in order to align a source narrative to narrative retellings elicited in a clinical setting. From these alignments, we automatically extract narrative recall scores which can then be used for diagnostic screening. The significant reduction in alignment error rate (AER) afforded by the graph-based method results in improved automatic scoring and diagnostic classification. The approach described here is general enough to be applied to almost any narrative recall scenario, and the reductions in AER achieved in this work attest to the potential utility of this graph-based method for enhancing multilingual word alignment and alignment of comparable corpora for more standard NLP tasks.

1 Introduction

Much of the work in biomedical natural language processing has focused on mining information from electronic health records, clinical notes, and medical literature, but NLP is also very well suited for analyzing patient language data, in terms of both content and linguistic features, for neurological evaluation. NLP-driven analysis of clinical language data has been used to assess language development (Sagae et al., 2005), language impairment (Gabani et al., 2009) and cognitive status (Roark et al., 2007; Roark et al., 2011). These approaches rely on the extraction of syntactic features from spoken language transcripts in order to identify characteristics of language use associated with a particular disorder. In this paper, rather than focusing on linguistic features, we instead propose an NLP-based method for automating the standard manual method for scoring the Wechsler Logical Memory (WLM) subtest of the Wechsler Memory Scale (Wechsler, 1997) with the eventual goal of developing a screening tool for Mild Cognitive Impairment (MCI), the earliest observable precursor to dementia. During standard administration of the WLM, the examiner reads a brief narrative to the subject, who then retells the story to the examiner, once immediately upon hearing the story and a second time after a 30-minute delay. The examiner scores the retelling in real time by counting the number of recalled *story elements*, each of which corresponds to a word or short phrase in the source narrative. Our method for automatically extracting the score from a retelling relies on an alignment between substrings in the retelling and substrings in the original narrative. The scores thus extracted can then be used for diagnostic classification.

Previous approaches to alignment-based narrative analysis (Prud'hommeaux and Roark, 2011a; Prud'hommeaux and Roark, 2011b) have relied exclusively on modified versions of standard word alignment algorithms typically applied to large bilingual parallel corpora for building machine translation models (Liang et al., 2006; Och et al., 2000). Scores extracted from the alignments produced using these algorithms achieved fairly high classifi-

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 1–10,
Montréal, Canada, June 8, 2012.

cation accuracy, but the somewhat weak alignment quality limited performance. In this paper, we compare these word alignment approaches to a new approach that uses traditionally-derived word alignments between retellings as the input for graph-based exploration of the alignment space in order to improve alignment accuracy. Using both earlier approaches and our novel method for word alignment, we then evaluate the accuracy of automated scoring and diagnostic classification for MCI.

Although the alignment error rates for our data might be considered high in the context of building phrase tables for machine translation, the alignments produced using the graph-based method are remarkably accurate given the small size of our training corpus. In addition, these more accurate alignments lead to gains in scoring accuracy and to classification performance approaching that of manually derived scores. This method for word alignment and score extraction is general enough to be easily adapted to other tests used in neuropsychological evaluation, including not only those related to narrative recall, such as the NEPSY Narrative Memory subtest (Korkman et al., 1998) but also picture description tasks, such as the Cookie Theft picture description task of the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001) or the Renfrew Bus Story (Glasgow and Cowley, 1994). In addition, this technique has the potential to improve word alignment for more general NLP tasks that rely on small corpora, such as multilingual word alignment or word alignment of comparable corpora.

2 Background

The act of retelling or producing a narrative taps into a wide array of cognitive functions, not only memory but also language comprehension, language production, executive function, and theory of mind. The inability to coherently produce or recall a narrative is therefore associated with many different cognitive and developmental disorders, including dementia, autism (Tager-Flusberg, 1995), and language impairment (Dodwell and Bavin, 2008; Botting, 2002). Narrative tasks are widely used in neuropsychological assessment, and many commonly used instruments and diagnostic protocols include a task involving narrative recall or production (Korkman et al., 1998; Wechsler, 1997; Lord et al., 2002).

In this paper, we focus on evaluating narrative recall within the context of Mild Cognitive Impairment (MCI), the earliest clinically significant precursor of dementia. The cognitive and memory problems associated with MCI do not necessarily interfere with daily living activities (Ritchie and Touchon, 2000) and can therefore be difficult to diagnose using standard dementia screening tools, such as the Mini-Mental State Exam (Folstein et al., 1975). A definitive diagnosis of MCI requires an extensive interview with the patient and a family member or caregiver. Because of the effort required for diagnosis and the insensitivity of the standard screening tools, MCI frequently goes undiagnosed, delaying the introduction of appropriate treatment and remediation. Early and unobtrusive detection will become increasingly important as the elderly population grows and as research advances in delaying and potentially stopping the progression of MCI into moderate and severe dementia.

Narrative recall tasks, such as the test used in research presented here, the Wechsler Logical Memory subtest (WLM), are often used in conjunction with other cognitive measures in attempts to identify MCI and dementia. Multiple studies have demonstrated a significant difference in performance on the WLM between subjects with MCI and typically aging controls, particularly in combination with tests of verbal fluency and memory (Storandt and Hill, 1989; Peterson et al., 1999; Nordlund et al., 2005). The WLM can also serve as a cognitive indicator of physiological characteristics associated with symptomatic Alzheimers disease, even in the absence of previously reported dementia (Schmitt et al., 2000; Bennett et al., 2006).

Some previous work on automated analysis of the WLM has focused on using the retellings as a source of linguistic data for extracting syntactic and phonetic features that can distinguish subjects with MCI from typically aging controls (Roark et al., 2011). There has been some work on automating scoring of other narrative recall tasks using unigram overlap (Hakkani-Tur et al., 2010), but Dunn et al. (2002) are among the only researchers to apply automated methods to scoring the WLM for the purpose of identifying dementia, using latent semantic analysis to measure the semantic distance between a retelling

Dx	n	Age	Education
MCI	72	88.7	14.9 yr
Non-MCI	163	87.3	15.1 yr

Table 1: Subject demographic data.

and the source narrative. Although scoring automation is not typically used in a clinical setting, the objectivity offered by automated measures is particularly important for tests like the WLM, which are often administered by practitioners working in a community setting and serving a diverse population.

Researchers working on NLP tasks such as paraphrase extraction (Barzilay and McKeown, 2001), word-sense disambiguation (Diab and Resnik, 2002), and bilingual lexicon induction (Sahlgren and Karlgren, 2005), often rely on aligned parallel or comparable corpora. Recasting the automated scoring of a neuropsychological test as another NLP task involving the analysis of parallel texts, however, is a relatively new idea. We hope that the methods presented here will both highlight the flexibility of techniques originally developed for standard NLP tasks and attract attention to the wide variety of biomedical data sources and potential clinical applications for these techniques.

3 Data

3.1 Subjects

The data examined in this study was collected from participants in a longitudinal study on brain aging at the Layton Aging and Alzheimers Disease Center at the Oregon Health and Science University (OHSU), including 72 subjects with MCI and 163 typically aging seniors roughly matched for age and years of education. Table 1 shows the mean age and mean years of education for the two diagnostic groups. There were no significant between-group differences in either measure.

Following (Shankle et al., 2005), we assign a diagnosis of MCI according to the Clinical Dementia Rating (CDR) (Morris, 1993). A CDR of 0.5 corresponds to MCI (Ritchie and Touchon, 2000), while a CDR of zero indicates the absence of MCI or any dementia. The CDR is measured via the Neurobehavioral Cognitive Status Examination (Kiernan et al., 1987) and a semi-structured interview with the patient and a family member or caregiver that allows the examiner to assess the subject in several key areas of cognitive function, such as memory, orientation, problem solving, and personal care. The CDR has high inter-annotator reliability (Morris, 1993) when conducted by trained experts. It is crucial to note that the calculation of CDR is completely independent of the neuropsychological test investigated in this paper, the Wechsler Logical Memory subtest of the Wechsler Memory Scale. We refer readers to the above cited papers for a further details.

3.2 Wechsler Logical Memory Test

The Wechsler Logical Memory subtest (WLM) is part of the Wechsler Memory Scale (Wechsler, 1997), a diagnostic instrument used to assess memory and cognition in adults. In the WLM, the subject listens to the examiner read a brief narrative, shown in Figure 1. The subject then retells the narrative to the examiner twice: once immediately upon hearing it (Logical Memory I, LM-I) and again after a 30-minute delay (Logical Memory II, LM-II). The narrative is divided into 25 *story elements*. In Figure 1, the boundaries between story elements are denoted by slashes. The examiner notes in real time which story elements the subject uses. The score that is reported under standard administration of the task is a summary score, which is simply the raw number of story elements recalled. Story elements do not need to be recalled verbatim or in the correct temporal order. The published scoring guidelines describe the permissible substitutions for each story element. The first story element, *Anna*, can be replaced in the retelling with *Annie* or *Ann*, while the 16th story element, *fifty-six dollars*, can be replaced with any number of dollars between fifty and sixty.

An example LM-I retelling is shown in Figure 2. According to the published scoring guidelines, this retelling receives a score of 12, since it contains the following 12 elements: *Anna, employed, Boston, as a cook, was robbed of, she had four, small children, reported, station, touched by the woman's story, took up a collection,* and *for her*.

3.3 Word alignment data

The Wechsler Logical Memory immediate and delayed retellings for all of the 235 experimental subjects were transcribed at the word level. We sup-

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

Figure 1: Text of WLM narrative segmented into 25 story elements.

> 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: Sample retelling of the Wechsler narrative.

plemented the data collected from our experimental subjects with transcriptions of retellings from 26 additional individuals whose diagnosis had not been confirmed at the time of publication or who did not meet the eligibility criteria for this study. 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 produced word-level alignments between each retelling and the source narrative presented in Figure 1.

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. Since 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 roughly 500-line source-to-retelling corpus consisting of the source narrative on one side and each retelling on the other.
- **Corpus 2:** A roughly 250,000-line pairwise retelling-to-retelling corpus, consisting of every possible pairwise combination of retellings.
- **Corpus 3:** A roughly 900-line word identity corpus, consisting of every word that appears in every retelling and the source narrative.

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 entirely unsupervised. 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 retellings from the 26 individuals who were not experimental subjects as a development set for tuning the various parameters of our system, which is described below.

4 Word Alignment

4.1 Baseline alignment

We begin by building two word alignment models using the Berkeley aligner (Liang et al., 2006), a state-of-the-art word alignment package that relies on IBM mixture 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 and because it offers functionality for testing an existing model on new data and for outputting posterior probabilities. The smaller of our two Berkeley-generated models is trained on Corpus 1 (the source-to-retelling parallel corpus described above) 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 subjects. 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.

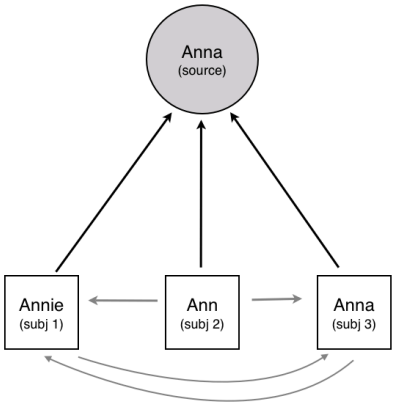

Figure 3: Depiction of word graph.

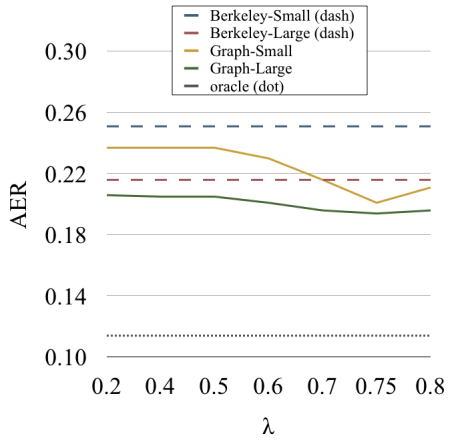

Figure 4: Changes in AER as λ increases.

The first two rows of Table 2 show the precision, recall, F-measure, and alignment error rate (AER) (Och and Ney, 2003) for these two Berkeley aligner models. We note that although AER for the larger model is lower, the time required to train the model is significantly larger. The alignments generated by the Berkeley aligner serve not only as a baseline for comparison but also as a springboard for the novel graph-based method of alignment we will now discuss.

4.2 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 various NLP tasks in information extraction and retrieval, including webpage ranking (PageRank (Page et al., 1999)) and extractive summarization (LexRank (Erkan and Radev, 2004; Otterbacher et al., 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 likelihood of a random walk through the graph starting at a particular node and ending at another node provides information about the relationship between those two nodes and the importance of the starting node.

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, as depicted in Figure 3. Starting at a particular node (i.e., a word in one of the retellings), our algorithm can either walk from that node to another node in the graph or to a word in the source narrative. At each step in the walk, there is a set probability λ that determines the likelihood of transitioning to another retelling word versus a word in the source narrative. When transitioning to a retelling word, the destination word is chosen according to the posterior probability assigned by the Berkeley aligner to that alignment. When the walk arrives at a source narrative word, that word is the new proposed alignment for the starting word.

For each word in each retelling, we perform 1000 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.

We build two graphs on which to carry out these random walks: one graph is built using the alignments generated by the smaller Berkeley alignment model, and the other is built from the alignments generated by the larger Berkeley alignment model. Alignments with posterior probabilities of 0.5 or greater are included as edges within the graph, since this is the default posterior threshold used by the Berkeley aligner. The value of λ, the probability of walking to a retelling word node rather than a source word, is tuned to the development set of retellings,

Model	P	R	F	AER
Berkeley-Small	72.1	79.6	75.6	24.5
Berkeley-Large	78.6	80.5	79.5	20.5
Graph-Small	77.9	81.2	79.5	20.6
Graph-Large	85.4	76.9	81.0	18.9

Table 2: Aligner performance comparison.

discussed in Section 3.3. Figure 4 shows how AER varies according to the value of λ for the two graph-based approaches.

Each of these four alignment models produces, for each retelling, a set of word pairs containing one word from the original narrative and one word from the retelling. The manual gold alignments for the 235 experimental subjects were evaluated against the alignments produced by each of the four models. Table 2 shows the accuracy of word alignment using these two graph-based models in terms of precision, accuracy, F-measure, and alignment error rate, alongside the same measures for the two Berkeley models. We see that each of the graph-based models outperforms the Berkeley model of the same size. The performance of the small graph-based model is especially remarkable since it an AER comparable to the large Berkeley model while requiring significantly fewer computing resources. The difference in processing time between the two approaches was especially remarkable: the graph-based model completed in only a few minutes, while the large Berkeley model required 14 hours of training.

Figures 5 and 6 show the results of aligning the retelling presented in Figure 2 using the small Berkeley model and the large graph-based model, respectively. Comparing these two alignments, we see that the latter model yields more precise alignments with very little loss of recall, as is borne out by the overall statistics shown in Table 2.

5 Scoring

The published scoring guidelines for the WLM specify the source words that compose each story element. Figure 7 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 extract scores from the alignments as follows: for each word in the original narrative, if

[A anna$_1$] [B thompson$_2$] [C of$_3$ south$_4$] [D boston$_5$] [E employed$_6$] [F as$_7$ a$_8$ cook$_9$] [G in$_{10}$ a$_{11}$ school$_{12}$] [H cafeteria$_{13}$] [I reported$_{14}$] [J at$_{15}$ the$_{16}$ police$_{17}$] [K station$_{18}$] [L that$_{19}$ she$_{20}$ had$_{21}$ been$_{22}$ held$_{23}$ up$_{24}$] [M on$_{25}$ state$_{26}$ street$_{27}$] [N the$_{28}$ night$_{29}$ before$_{30}$] [O and$_{31}$ robbed$_{32}$ of$_{33}$] [P fifty-six$_{34}$ dollars$_{35}$] [Q she$_{36}$ had$_{37}$ four$_{38}$] [R small$_{39}$ children$_{40}$] [S the$_{41}$ rent$_{42}$ was$_{43}$ due$_{44}$] [T and$_{45}$ they$_{46}$ had$_{47}$ n't$_{48}$ eaten$_{49}$] [U for$_{50}$ two$_{51}$ days$_{52}$] [V the$_{53}$ police$_{54}$] [W touched$_{55}$ by$_{56}$ the$_{57}$ woman's$_{58}$ story$_{59}$] [X took$_{60}$ up$_{61}$ a$_{62}$ collection$_{63}$] [Y for$_{64}$ her$_{65}$]

Figure 7: Text of Wechsler Logical Memory narrative with story-element labeled bracketing and word IDs.

anna(1) : A	robbed(32) : O	station(18) : K
thompson(2) : B	fifty-six(34) : P	took(60) : X
employed(6) : E	four(38) : Q	collection(63) : X
boston(5) : D	children(40) : R	for(64) : Y
cook(9) : F	reported(14) : I	her(65) : Y

Figure 8: Source content words from the alignment in Figure 6 with corresponding story element IDs.

that word is aligned to a word in the retelling, the story element that it is associated with is considered to be recalled. Figure 8 shows the story elements extracted from the word alignments in Figure 6.

When we convert alignments to scores in this way, any alignment can be mapped to an element, even an alignment between function words such as *the* and *of*, which would be unlikely to indicate that the story element had been recalled. To avoid such scoring errors, we disregard any word-alignment pair containing a source function word. 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.

The element-level scores induced from the four word alignments for all 235 experimental subjects were evaluated against the manual per-element scores. We report the precision, recall, and f-measure for all four alignment models in Table 3. In addition, 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-

ann(1) : anna(1)
worked(3) : employed(6)
in(4) : in(10)
boston(5) : boston(5)
as(6) : as(7)
a(7) : a(8)
cook(8) : cook(9)
and(9) : and(31)
robbed(12) : robbed(32)
of(13) : of(33)
dollars(15) : dollars(35)
is(16) : was(43)
that(17) : that(19)
and(19) : and(45)
she(20) : she(36)
had(21) : had(37)
four(22) : four(38)
children(23) : children(40)
reported(25) : reported(14)
at(26) : at(15)
the(27) : the(16)
some(28) : police(17)
station(31) : station(18)
made(37) : up(61)
made(37) : took(60)
a(38) : a(62)
collection(39) : collection(63)
for(40) : for(64)
her(41) : her(65)
so(42) : woman's(58)
she(44) : she(20)

Figure 5: Word alignment generated by the small Berkeley alignment model with retelling words italicized.

ann(1) : anna(1)
taylor(2) : thompson(2)
worked(3) : employed(6)
in(4) : in(10)
boston(5) : boston(5)
as(6) : as(7)
a(7) : a(8)
cook(8) : cook(9)
robbed(12) : robbed(32)
of(13) : of(33)
sixty-seven(14) : fifty-six(34)
dollars(15) : dollars(35)
she(20) : she(36)
had(21) : had(37)
four(22) : four(38)
children(23) : children(40)
reported(25) : reported(14)
at(26) : at(15)
the(27) : the(16)
station(31) : station(18)
made(37) : took(60)
a(38) : a(62)
collection(39) : collection(63)
for(40) : for(64)
her(41) : her(65)

Figure 6: Word alignment generated by the large graph-based model with retelling words italicized.

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

Table 3: Scoring accuracy results.

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 et al., 2003). As will be shown in the following section, scoring accuracy is very important for achieving high diagnostic classification accuracy, which is the ultimate goal of this work.

6 Diagnostic Classification

As discussed in Section 2, poor performance on the Wechsler Logical Memory test is associated with Mild Cognitive Impairment. We now use the scores we have extracted from the word alignments as features with a support vector machine (SVM) to perform diagnostic classification for distinguishing subjects with MCI from those without. For each of the 235 experimental subjects, we generate 2 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. In addition to the summary score, we also provide the SVM with a vector of 50 per-element scores: for each of the 25 element in each of the two retellings per subject, there is a vector element with the value of 0 if the element was not recalled, or 1 if the element was recalled. Since previous work has indicated that certain elements may be more powerful in their ability to predict the presence of MCI, we expect that giving the SVM these per-elements scores may improve classification performance. To train and test our classifiers, we use the WEKA API (Hall et al., 2009) and LibSVM (Chang and Lin, 2011), with a second-order polynomial kernel and default parameter settings.

We evaluate the performance of the SVMs using a leave-pair-out validation scheme (Cortes et al., 2007; Pahikkala et al., 2008). In the leave-pair-out technique, every pairing between a negative example and a positive example is tested using a classifier trained on all of the remaining examples. The resulting pairs of scores can be used to calculate the area under the receiver operating characteristic (ROC) curve (Egan, 1975), which is a plot of the false positive rate of a classifier against its true positive rate. The area under this curve (AUC) has a

Model	Summ. (s.d.)	Elem. (s.d.)
Manual Scores	73.3 (3.76)	81.3 (3.32)
Berkeley-Small	73.7 (3.74)	77.9 (3.52)
Berkeley-Big	75.1 (3.67)	79.2 (3.45)
Graph-Small	74.2 (3.71)	78.9 (3.47)
Graph-Big	74.8 (3.69)	78.6 (3.49)

Table 4: Classification accuracy results (AUC).

value of 0.5 when the classifier performs at chance and a value 1.0 when perfect classification accuracy is achieved.

Table 4 shows the classification results for the scores derived from the four alignment models along with the classification results using the examiner-assigned manual scores. 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, this very slight margin of difference may not justify its significantly greater computational requirements.

7 Conclusions and Future Work

The work presented here demonstrates the utility of adapting techniques drawn from a diverse set of NLP research areas to tasks in biomedicine. 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. Our novel graph-based approach to word alignment resulted in large reductions in alignment error rate. These reductions in error rate in turn led to human-level scoring accuracy and improved diagnostic classification.

As we have mentioned, the methods outlined here are general enough to be used for other episodic recall and description scenarios. Although the results are quite robust, several enhancements and improvements should be made before we apply the system to other tasks. First, although we were able to achieve decent word alignment accuracy, especially with our graph-based approach, many alignment errors remain. As shown in Figure 4, the graph-based alignment technique could potentially result in an AER of as low as 11%. We expect that our decision to select as a new alignment the most frequent source word over the distribution of source words at the end of 1000 walks could be improved, since it does not allow for one-to-many mappings. In addition, it would be worthwhile to experiment with several posterior thresholds, both during the decoding step of the Berkeley aligner and in the graph edges.

In order to produce a viable clinical screening tool, it is crucial that we incorporate speech recognition in the pipeline. Our very preliminary investigation into using ASR to generate transcripts for alignment seems promising and surprisingly robust to the problems that might be expected when working with noisy audio. In our future work, we also plan to examine longitudinal data for individual subjects to see whether our techniques can detect subtle differences in recall and coherence between a recent retelling and a series of earlier baseline retellings. Since the metric commonly used to quantify the progression of dementia, the Clinical Dementia Rating, relies on observed changes in cognitive function over time, longitudinal analysis of performance on the Wechsler Logical Memory task may be the most promising application for our research.

References

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceeding of ACL*.

D.A. Bennett, 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*, 66:1837–844.

Nicola Botting. 2002. Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy*, 18(1).

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statis-

tical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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

Corinna Cortes, Mehryar Mohri, and Ashish Rastogi. 2007. An alternative ranking problem for search engines. In *Proceedings of WEA2007, LNCS 4525*, pages 1–21. Springer-Verlag.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL*.

Kristy Dodwell and Edith L. Bavin. 2008. Children with specific language impair ment: an investigation of their narratives and memory. *International Journal of Language and Communication Disorders*, 43(2):201–218.

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

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

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479.

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

Keyur Gabani, 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 NAACL-HLT*, pages 46–55.

Cheryl Glasgow and Judy Cowley. 1994. *Renfrew Bus Story test - North American Edition*. Centreville School.

H Goodglass, E Kaplan, and B Barresi. 2001. *Boston Diagnostic Aphasia Examination. 3rd ed.* Pro-Ed.

Dilek Hakkani-Tur, Dimitra Vergyri, and Gokhan Tur. 2010. Speech-based automated cognitive status assessment. In *Proceedings of Interspeech*, pages 258–261.

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

David K. Johnson, 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.

R.J. Kiernan, J. Mueller, J.W. Langston, and C. Van Dyke. 1987. The neurobehavioral cognitive status examination, a brief but differentiated approach to cognitive assessment. *Annals of Internal Medicine*, 107:481–485.

Marit Korkman, Ursula Kirk, and Sally Kemp. 1998. *NEPSY: A developmental neuropsychological assessment*. The Psychological Corporation.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT NAACL*.

Catherine Lord, Michael Rutter, Pamela DiLavore, and Susan Risi. 2002. *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services.

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

A Nordlund, 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.

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

Franz Josef Och, Christoph Tillmann, , and Hermann Ney. 2000. Improved alignment models for statistical machine translation. In *Proceedings of ACL*, pages 440–447.

Jahna Otterbacher, Günes Erkan, and Dragomir R. Radev. 2009. Biased lexrank: Passage retrieval using random walks with question-based priors. *Inf. Process. Manage.*, 45(1):42–54.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November. Previous number = SIDL-WP-1999-0120.

Tapio Pahikkala, Antti Airola, Jorma Boberg, and Tapio Salakoski. 2008. Exact and efficient leave-pair-out cross-validation for ranking RLS. In *Proceedings of AKRR 2008*, pages 1–8.

Ronald Peterson, 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.

Emily T. Prud'hommeaux and Brian Roark. 2011a. Alignment of spoken narratives for automated neuropsychological assessment. In *Proceedings of ASRU*.

Emily T. Prud'hommeaux and Brian Roark. 2011b. Extraction of narrative recall patterns for neuropsychological assessment. In *Proceedings of Interspeech*.

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

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the ACL 2007 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 1–8.

Brian Roark, 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.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of ACL*, pages 197–204.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3).

F.A. Schmitt, 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.

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

Martha Storandt and Robert Hill. 1989. Very mild senile dementia of the alzheimers type: Ii psychometric test performance. *Archives of Neurology*, 46:383–386.

Helen Tager-Flusberg. 1995. Once upon a ribbit: Stories narrated by autistic children. *British journal of developmental psychology*, 13(1):45–59.

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