Extraction of narrative recall patterns for neuropsychological assessment

Emily T. Prud'hommeaux, Brian Roark

Center for Spoken Language Understanding
Oregon Health and Science University, Portland, Oregon, USA
emily@cslu.ogi.edu, roark@cslu.ogi.edu

Abstract

Poor narrative memory is associated with a variety of neurodegenerative and developmental disorders, such as autism and Alzheimer's related dementia. Hence, narrative recall tasks are included in most standard neurological examinations. In this paper, we explore methods for automatically assessing the quality of retellings via alignment to the original narrative. Word alignments serve both to automate manual scoring and to derive other features related to narrative coherence that can be used for diagnostic classification. Despite relatively high word alignment error rates, the automatic alignments provide sufficient information to achieve nearly as accurate diagnostic classification as manual scores. Furthermore, additional features that become available with alignment provide utility in classifying subject groups. While the additional features we explore here did not provide additive gains in accuracy, they point the way to the development of many potentially useful features in this domain.
Index Terms: spoken language health applications, classification, alignment.

1. Introduction

Narrative recall tasks, in which a subject must listen to and then retell a story, provide a way to elicit semi-structured spontaneous language data ideal for analysis with natural language processing techniques. A number of widely used neuropsychological instruments include a narrative recall or production task [1, 2, 3], and narrative recall ability is a good predictor of a variety of cognitive and developmental problems such as language impairment [4, 5], autism [6, 7], and dementia [8].

Mild Cognitive Impairment (MCI), the earliest observable stage of dementia, is associated with cognitive symptoms that do not always interfere with daily living activities [9] and can thus be difficult to detect using the common screening tests for dementia. An extensive interview with the patient and caregiver is usually necessary for accurate diagnosis. MCI often goes undiagnosed, which results in significant delays in treatment [10].

In this paper, we present a new method for automatically analyzing narrative retellings in order to provide additional information to be used in the diagnosis of MCI. We explore using word-level alignments, both manually and automatically derived, for assessing the fidelity of a narrative retelling to the original narrative, as measured by the number of key story elements used in the retelling. We then present some previously unexplored diagnostic language features that can be obtained for free as a by-product of word alignment and show that these features can be used within a machine learning framework for diagnostic classification of MCI. While we did not achieve additive gains by combining all of the features, their individual utility bodes well for a future feature engineering effort.

Our work focuses on the Logical Memory narrative recall subtest of the Wechsler Memory Scale [2], a widely used instrument for evaluating memory function in adults. The techniques described here, however, have the potential to be easily adapted to a number of narrative production scenarios, including narrative memory tests designed for detecting language impairment and tasks testing narrative production from a silent movie or wordless picture book.

2. Related Work

Most previous work in using natural language processing (NLP) for analysis of spoken language or narratives for diagnostic classification has focused not on evaluating coherence but rather on using the the language samples as a data source from which to extract speech and language features. Several studies of language and dementia have examined spontaneous speech for particular speech characteristics (e.g., pauses) or language complexity [11, 12, 13]. Research on diagnosing language disorders in children has investigated automating standardized annotation of spontaneous speech for assessing language development [14] and using language modeling to address problems in evaluating bilingual children using common diagnostic instruments [15].

A few recent studies have examined features in narratives similar to those investigated in non-narrative spontaneous speech. In their study of specific language impairment (SLI), Gabbani et al. [16] found potential for improved identification of SLI using measures of morphology, fluency, vocabulary, and linguistic productivity derived automatically from children's recalled narratives and spontaneous narratives. Analyzing data from the Wechsler Logical Memory task, Roark et al. [17] investigated using measures of syntactic complexity to discriminate between individuals with and without MCI.

Although work in the neuropsychology community has demonstrated the clinical significance of performance on narrative recall tasks, the quality of the retellings relative to the original narrative was not considered in the above studies. In contrast, the work presented here focuses primarily on the accuracy and coherence of the retelling itself.

3. Data

3.1. Subjects

Seventy-two subjects with MCI and fifty-two healthy seniors, roughly matched for age and years of education, were selected from a large pool of subjects taking part in a study of brain aging at the Layton Aging and Alzheimer's Disease Center at the Oregon Health and Science University (OHSU). MCI group membership was determined using the Clinical Dementia Rating (CDR) [18], following previous work in this area [19]. MCI was defined as a CDR of 0.5, while the absence of MCI was defined as a CDR of zero. The CDR was chosen here because it has strong expert inter-annotator reliability [18] and is as-

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

Figure 1: Text of Wechsler Logical Memory narrative

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.

signed independently of the Wechsler Logical Memory test investigated in this paper. We refer readers to these cited papers for a more detailed discussion of the CDR.

3.2. Wechsler Logical Memory Test

In the Wechsler Logical Memory task, one of several components of the Wechsler Memory Scale [2], the subject listens to a brief story, shown in Figure 1 and must then retell the story first immediately after hearing it (Logical Memory I, LM-I) and a second time after a 30-minute delay (Logical Memory II, LM-II). For scoring purposes, the text is segmented into 25 *story elements*, denoted in Figure 1 with slashes. During examination, the examiner notes which story elements were recalled and reports a summary score, which is simply the total number of story elements used. Elements can be paraphrased and they do not need to be in order. The scoring guidelines describe the permissible substitutions, such as *Ann* for the element, *Anna*.

Figure 2 shows an example LM-I retelling, which includes the 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*, yielding a summary score of 12 according to published guidelines for manual scoring. Note the change of ordering of some of the story elements (*worked* comes before *Boston* rather than after); the off-topic aside ('*Is that right?*); and embellishments (*so that she can feed the children*).

3.3. Data

Audio recordings of the Wechsler Logical Memory retellings for the 124 subjects were transcribed, tokenized, and normalized. A subset of the retellings were then hand-aligned to the original narrative at the word level for alignment evaluation purposes. Given the presence of frequent omissions and embellishments, many words in the retellings were not aligned directly to any word in the original narrative and vice versa.

Every word or word sequence in every retelling matching one of the story elements was identified manually according to the scoring guidelines. We also identified words and phrases that were logical substitutes that would not have been acceptable for scoring purposes, such as *Taylor* instead of *Thompson*. Figure 3 provides the list of words and phrases identified for the retelling in Figure 2, with logically acceptable substitutes that would not have been clinically acceptable in italics.

In our training data for word alignment we included these

Original Narrative	**Retelling**
anna	ann
thompson	*taylor*
employed	worked
boston	in boston
as a cook	as a cook
and robbed of	she was robbed of
fifty-six dollars	*sixty-seven dollars*
she had four	she had four
reported	reported
station	station
small children	children
the police	*the fellow*
touched by the woman's story	was sympathetic
took up a collection	and made a collection
for her	for her

Figure 3: Phrase-aligned sample retelling.

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

phrase-to-phrase parallel alignments, as well as the corpus of full retellings to the full original narrative and a list of word identity parallel alignments. Each speaker's retellings were aligned using a model built on these three kinds of parallel alignments derived from the retellings of the 123 subjects not being tested.

With the explicit boundaries between story elements provided in the scoring guidelines, we know precisely which words in the original narrative each story element contains. Figure 4 shows the original narrative labeled with element IDs ($A - Y$) and word IDs ($1 - 65$). Story element F, for instance, is composed of words 7, 8, and 9, *as a cook*. This information enables us to convert word alignments into story element scores for automatic scoring.

In the next sections we briefly discuss the performance of existing aligners on our data in terms of both alignment error rate and scoring accuracy, as well as the utility of automated scores and other alignment-derived features for subject classification.

4. Alignment and Automated Scoring

Each retelling was aligned to the original narrative using the Berkeley aligner [20] with default parameters. A subset of these alignments were evaluated against the manually generated word alignments available for that subset. Figure 5 shows the results of aligning the retelling presented in Figure 2 using the Berkeley aligner.

In order to automatically score the retellings using these alignments, we determined for each retelling the identities of

ann(1) : anna(1)
taylor(2) : thompson(2)
worked(3) : employed(6)
in(4) : null
boston(5) : boston(5)
as(6) : as(7)
a(7) : a(8)
cook(8) : cook(9)
and(9) : and(31)
she(10) : null
was(11) : null
robbed(12) : robbed(32)
of(13) : of(33)
sixty-seven(14) : null
dollars(15) : fifty-six(34)
is(16) : null
that(17) : null
right(18) : due(44)
and(19) : and(45)
she(20) : she(36)
had(21) : had(37)
four(22) : four(38)
children(23) : small(39)
and(24) : null
reported(25) : reported(14)
at(26) : at(15)
the(27) : the(16)
some(28) : on(25)
kind(29) : night(29)
of(30) : null
station(31) : station(18)
the(32) : null
fellow(33) : null
was(34) : null
sympathetic(35) : story(59)
and(36) : null
made(37) : took(60)
a(38) : a(62)
collection(39) : collection(63)
for(40) : for(64)
her(41) : her(65)
so(42) : null
that(43) : that(19)
she(44) : she(20)
can(45) : been(22)
feed(46) : eaten(49)
her(47) : null
children(48) : null

Figure 5: Word-alignment generated by the Berkeley aligner with retelling words italicized.

anna(1) : A
thompson(2) : B
employed(6) : E
boston(5) : D
cook(9) : F
robbed(32) : O
fifty-six(34) : P
due(44) : S
four(38) : Q
small(39) : R
reported(14) : I
night(29) : N
station(18) : K
story(59) : W
took(60) : X
collection(63) : X
for(64) : Y
her(65) : Y
eaten(49) : T

Figure 6: Words from original narrative from the alignment in Fig. 5, excluding function words, with story element IDs.

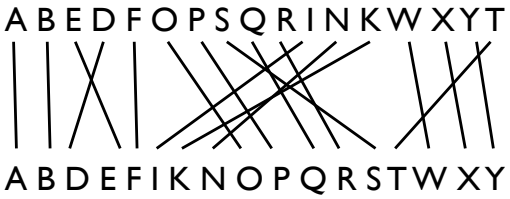

Figure 7: Element ordering with crossing links.

the story elements recalled using the following simple procedure. Recall that every word in the original narrative is associated with a particular story element ID, as shown in Figure 4. We consider each word in the original narrative in turn; if the word is aligned to a word in the retelling, the story element ID that it is associated with is considered to be correctly recalled. In Figure 6, we see the list of story elements induced from the word alignments in Figure 5.

The story element identities induced from the automatically derived word alignments were evaluated against the manually derived per-element scores, resulting in an F1 measure of 0.79, with recall (0.97) higher than precision (0.67). Although the word alignment error rate (AER) for the Berkeley aligner was quite high (31%) compared to the AER typically reported for the Berkeley aligner when trained on a large amount of data (4.9% AER on the Hansards English-French task) [20], the scoring accuracy is remarkably robust, demonstrating the utility of applying existing algorithms to this novel task.

5. Classification

5.1. Method

We now move on to using these scores and other information extracted from the word alignments as features within a support vector machine (SVM) to perform classification of subjects into the two diagnostic groups: individuals with and without mild cognitive impairment. We collected 3 pairs of features per subject. First, from the scores extracted as described in Section 4, we counted the number of elements recalled for each retelling, generating two summary scores per subject, one each for LM-I and LM-II, ranging from 0 (when no items were recalled) to 25 (when all items were recalled).

Next we calculated the amount of off-topic or irrelevant content in each subject's two retellings, which we represented as the ratio of unaligned words to total words used in a retelling. In Figure 5, we see that a number of words from the retelling are aligned to null. These words aligned to null compose the set of unaligned words used for this measure of irrelevance.

Finally, we derived a feature to measure the quality of the ordering of story elements in each retelling relative to the order in which they appeared in the original narrative. Story elements were first extracted as described in Section 4 and shown in Figure 6. In a sequence of N retold elements, for each element, e_i, we then counted the number of following elements, $e_{i+1} \ldots e_N$ that were incorrectly ordered relative to e_i. This measure captures in a simple way how far away each element is from its position in a correctly ordered sequence. This measure can be visualized, as in Figure 7, as the number of crossing links between the observed ordering and the correct ordering. This measure was then compared to the average measure of 100 random orderings of the sequence. The feature reported here is the ratio between the observed ordering measure to the average ordering measure over these 100 random orderings.

LibSVM [21], the SVM class included in the Waikato Environment for Knowledge Analysis [22], was used to train support vector machine classifiers with default parameter settings. Each feature value was scaled to range between 0 and 1 according to the minimum and maximum values for the feature in the training data.

To evaluate our models, we chose to use a leave-pair-out validation scheme [23, 24] in which every possible pair consisting of a negative example and a positive example is tested against an SVM classifier trained on the remaining 122 subjects. The resulting pairs of scores can be used to calculate the area under the receiver operating characteristic (ROC) curve [25], a commonly used estimate of classification accuracy. The area under this curve (AUC) can range from 0.5, when the classifier performs no better than chance, to 1.0, when the classifier has perfect accuracy.

5.2. Results

We first performed classification using each of the above three features individually and then tested combinations of features. We see in Table 1 that classification accuracy was quite high overall and that the classification accuracy of the automatically derived scores and features individually was usually quite comparable to that of the manually derived scores and alignments.

Feature Set	Automatic	Manual
Summary scores	0.795	0.822
Irrelevant content	0.492	0.704
Ordering	0.706	0.759
Irrelevant+Sum	0.799	0.817
Order+Sum	0.79	0.819
Irrelevant+Order	0.678	0.763
All	0.798	0.816

Table 1: Classification accuracy (AUC) results.

The summary scores provide the most discriminative power in both cases. The irrelevant content feature was not effective when automatically extracted, likely due to the high recall at the expense of precision achieved by the Berkeley aligner, which points to the need for improved automatic alignment techniques. Although these features achieved decent AUC individually, there was no additive gain with combination.

6. Discussion and Future Directions

The results presented here demonstrate that existing word-alignments techniques can be used to produce alignments sufficient for the extraction of narrative coherence features with significant diagnostic classification power. We also observe that the features derived from automated alignments yield classification accuracy comparable to that produced using manual alignments. This is an entirely novel application of existing techniques that offers substantial clinical utility.

Given these promising results, we plan to use our system with narrative recall data from the NEPSY, a neurodevelopmental protocol for detecting language impairment in children. We also will adapt our technique to structured narrative generation tasks, in which a subject tells a story to explain a silent movie or wordless picture book.

Although we were able to achieve high automatic scoring and classification accuracy by relying on existing techniques for word alignment, these techniques were designed to be used with an entirely different kind of data, specifically, large, multilingual corpora, in which each sentence in the source language has an equivalent in the the target language. Our corpora are very small and contain many repetitions, reorderings, omissions (when the subject fails to recall one of the story elements), and insertions (when the subject produces off-topic phrases or embellishments). In addition, our data is monolingual, and the techniques used for word alignment for machine translation make no use of such potentially helpful features as word identity during alignment. These issues will best be addressed by developing new alignment techniques tailored to this specific type of alignment task.

7. References

[1] M. Korkman, U. Kirk, and S. Kemp, *NEPSY: A developmental neuropsychological assessment*. The Psychological Corporation, 1998.

[2] D. Wechsler, *Wechsler Memory Scale - Third Edition Manual*. The Psychological Corporation, 1997.

[3] C. Lord, M. Rutter, P. DiLavore, and S. Risi, *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services, 2002.

[4] K. Dodwell and E. L. Bavin, "Children with specific language impair ment: an investigation of their narratives and memory," *International Journal of Language and Communication Disorders*, vol. 43, no. 2, pp. 201–218, 2008.

[5] N. Botting, "Narrative as a tool for the assessment of linguistic and pragmatic impairments," *Child Language Teaching and Therapy*, vol. 18, no. 1, 2002.

[6] J. J. Diehl, L. Bennetto, and E. C. Young, "Story recall and narrative coherence of high-functioning children with autism spectrum disorders," *Journal of Abnormal Child Psychology*, vol. 34, no. 1, pp. 87–102, 2006.

[7] H. Tager-Flusberg, "Once upon a ribbit: Stories narrated by autistic children," *British journal of developmental psychology*, vol. 13, no. 1, pp. 45–59, 1995.

[8] G. Gomez and D. A. White, "Using verbal fluency to detect very mild dementia of the alzheimer type," *Archives of Clinical Neuropsychology*, vol. 21, pp. 771–775, 2006.

[9] K. Ritchie and J. Touchon, "Mild cognitive impairment: Conceptual basis and current nosological status," *Lancet*, vol. 355, pp. 225–228, 2000.

[10] L. Boise, M. Neal, and J. Kaye, "Dementia assessment in primary care: Results from a study in three managed care systems," *Journal of Gerontology*, vol. 59, no. 6, 2004.

[11] S. Kemper, E. LaBarge, F. Ferraro, H. Cheung, H. Cheung, and M. Storandt, "On the preservation of syntax in Alzheimer's disease," *Archives of Neurology*, vol. 50, pp. 81–86, 1993.

[12] K. Lyons, S. Kemper, E. LaBarge, F. Ferraro, D. Balota, and M. Storandt, "Oral language and Alzheimer's disease: A reduction in syntactic complexity," *Aging and Cognition*, vol. 1, no. 4, pp. 271–281, 1994.

[13] S. Singh, R. Bucks, and J. Cuerden, "Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech," *Aphasiology*, vol. 15, no. 6, pp. 571–584, 2001.

[14] K. Sagae, A. Lavie, and B. MacWhinney, "Automatic measurement of syntactic development in child language," in *Proceedings of ACL*, 2005, pp. 197–204.

[15] T. Solorio and Y. Liu, "Using language models to identify language impairment in spanish-english bilingual children," in *Proceedings of the ACL Workshop on (BioNLP)*, 2008, pp. 116–117.

[16] K. Gabani, M. Sherman, T. Solorio, and Y. Liu, "A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children," in *Proceedings of NAACL-HLT*, 2009, pp. 46–55.

[17] B. Roark, M. Mitchell, and K. Hollingshead, "Syntactic complexity measures for detecting mild cognitive impairment," in *Proceedings of the ACL 2007 Workshop on Biomedical Natural Language Processing (BioNLP)*, 2007, pp. 1–8.

[18] J. Morris, "The clinical dementia rating (CDR): Current version and scoring rules," *Neurology*, vol. 43, pp. 2412–2414, 1993.

[19] W. R. Shankle, A. K. Romney, J. Hara, D. Fortier, M. B. Dick, J. M. Chen, T. Chan, and X. Sun, "Methods to improve the detection of mild cognitive impairment," *Proceedings of the National Academy of Sciences*, vol. 102, no. 13, pp. 4919–4924, 2005.

[20] P. Liang, B. Taskar, and D. Klein, "Alignment by agreement," in *Proceedings of the Human Language Technology Conference of the NAACL*, 2006.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[23] C. Cortes, M. Mohri, and A. Rastogi, "An alternative ranking problem for search engines," in *Proceedings of WEA2007, LNCS 4525*. Springer-Verlag, 2007, pp. 1–21.

[24] T. Pahikkala, A. Airola, J. Boberg, and T. Salakoski, "Exact and efficient leave-pair-out cross-validation for ranking RLS," in *Proceedings of AKRR 2008*, 2008, pp. 1–8.

[25] J. Egan, *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.