Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment

Maider Lehr[†], **Izhak Shafran**[†], **Emily Prud'hommeaux**[°] **and Brian Roark**[†]
[†]Center for Spoken Language Understanding, Oregon Health & Science University
[°]Center for Language Sciences, University of Rochester
{maiderlehr,zakshafran,emilpx,roarkbr}@gmail.com

Abstract

Automatically assessing the fidelity of a retelling to the original narrative – a task of growing clinical importance – is challenging, given extensive paraphrasing during retelling along with cascading automatic speech recognition (ASR) errors. We present a word tagging approach using conditional random fields (CRFs) that allows a diversity of features to be considered during inference, including some capturing acoustic confusions encoded in word confusion networks. We evaluate the approach under several scenarios, including both supervised and unsupervised training, the latter achieved by training on the output of a baseline automatic word-alignment model. We also adapt the ASR models to the domain, and evaluate the impact of error rate on performance. We find strong robustness to ASR errors, even using just the 1-best system output. A hybrid approach making use of both automatic alignment and CRFs trained tagging models achieves the best performance, yielding strong improvements over using either approach alone.

1 Introduction

Narrative production tasks are an essential component of many standard neuropsychological test batteries. For example, narration of a wordless picture book is part of the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002) and retelling of previously narrated stories is part of both the Developmental Neuropsychological Assessment (NEPSY) (Korkman et al., 1998) and the Wechsler Logical Memory (WLM) test (Wechsler, 1997). Such tests also arise in reading comprehension, second language learning and other computer-based tutoring systems (Xie et al., 2012; Zhang et al., 2008).

The accuracy of automated scoring of a narrative retelling depends on correctly identifying which of the source narrative's propositions or events (what we will call 'story elements') have been included in the retelling. Speakers may choose to relate these elements using diverse words or phrases, and an automated method of identifying these elements needs to model the permissible variants and paraphrasings. In previous work (Lehr et al., 2012; Prud'hommeaux and Roark, 2012; Prud'hommeaux and Roark, 2011), we developed models based on automatic word-alignment methods, as described briefly in Section 3. Such alignments are learned in an unsupervised manner from a parallel corpus of manual or ASR transcripts of retellings and the original source narrative, much as in machine translation training.

Relying on manual transcripts to train the alignment models limits the ability of these methods to handle ASR errors. By instead training on ASR transcripts, these methods can automatically capture some regularities of lexical variants and their common realizations by the recognizer. Additionally, evidence of acoustic confusability is available in word lattice output from the recognizer, which can be exploited to yield more robust automatic scoring, particularly in high error-rate scenarios.

In this paper, we present and evaluate the use of word tagging models for this task, in contrast to just using automatic (unsupervised) word-alignment methods. The approach is general enough to al-

Proceedings of NAACL-HLT 2013, pages 211–220,
Atlanta, Georgia, 9–14 June 2013.

low tagging of word confusion networks derived from lattices, thus allowing us to explore the utility of such representations to achieve robustness. We present results under a range of experimental conditions, including: variously adapting the ASR models to the domain; using maximum entropy models rather than CRFs; differing tagsets (BIO versus IO); and with varying degrees of supervision. Finally, we demonstrate improved utility in terms of using the automatic scores to classify elderly individuals as having Mild Cognitive Impairment. Ultimately we find that hybrid approaches, making use of both word-alignment and tagging models, yield strong improvements over either used independently.

2 Wechsler Logical Memory (WLM) task

The Wechsler Logical Memory (WLM) task (Wechsler, 1997), a widely used subtest of a battery of neuropsychological tests used to assess memory function in adults, has been shown to be a good indicator of Mild Cognitive Impairment (MCI) (Storandt and Hill, 1989; Petersen et al., 1999; Wang and Zhou, 2002; Nordlund et al., 2005), the stage of cognitive decline that is often a precursor to dementia of the Alzheimer's type. In the WLM, the subject listens to the examiner read a brief narrative and then retells the narrative twice: immediately upon hearing it and after about 20 minutes. The examiner grades the subject's response by counting how many of the story elements the subject recalled.

An excerpt of the text read by the clinician while administering the WLM task is shown in Figure 1. The story elements in the text are delineated using slashes, 25 elements in all. An example retelling is shown in Figure 2 to illustrate how the retellings are scored. The clinical evaluation guidelines specify what lexical substitutions, if any, are allowed for each element. Some elements, such as *cafeteria* and *Thompson*, must be recalled verbatim. In other cases, subjects are given credit for variants, such as *Annie* for *Anna*, or paraphrasing of concepts such as *sympathetic* for *touched by the woman's story*. The example retelling received a score of 12, with one point for each of the recalled story elements: *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*.

> 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 / . . . / police / touched by the woman's story / took up a collection / for her.

Figure 1: Reference text and the set of 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 sympathetic and made a collection for her so that she can feed the children.

Figure 2: An example retelling with 12 recalled story elements.

3 Unsupervised generative automated scoring with word alignment

In previous work (Lehr et al., 2012; Prud'hommeaux and Roark, 2012; Prud'hommeaux and Roark, 2011), we developed a pipeline for automatically scoring narrative retellings for the WLM task. The utterances corresponding to a retelling were recognized using an ASR system. The story elements were identified from the 1-best ASR transcript using word alignments produced by the Berkeley aligner (Liang et al., 2006), an EM-based word alignment package developed to align parallel texts for machine translation. The word alignment model was estimated in an unsupervised manner from a parallel corpus consisting of source narrative and manual transcripts of retellings from a small set of training subjects, and from a pairwise parallel corpus of manual retelling transcripts.

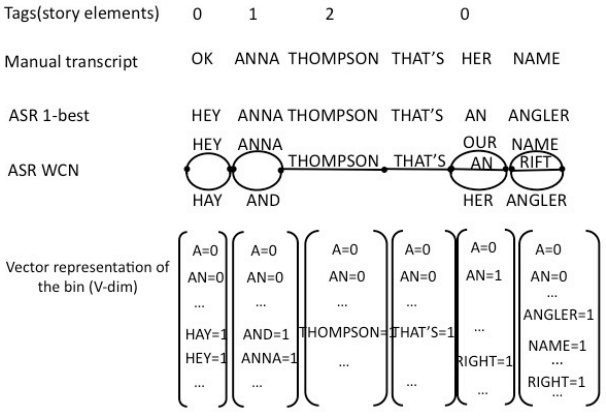
During inference or test, the ASR transcripts of the retellings were aligned using the estimated alignment model to the source narrative text. If a word in the retelling was mapped by the alignment model to a content word in the source narrative, the element associated with that content word was counted as correctly recalled in that retelling. Recall that the models were trained on unsupervised data so the aligned words may not always be permissible variants of the target elements. To alleviate such extraneous as well as unaligned words, the alignments below a threshold of posterior probability are discarded while decoding.

4 Supervised discriminative automated scoring with log-linear models

In this work, we frame the task of detecting story elements as a tagging task. Thus, our problem reduces to assigning a tag to each word position in the retelling, the tag indicating the story element that the word is associated with. In its simplest form, we have 26 tags: one for each of the 25 story elements indicating the word is 'in' that element (e.g., I15); and one for 'outside' of any story element ('O'). By tagging word positions, we are framing the problem in a general enough way to allow tagging of word confusion networks (Mangu et al., 2000), which encode word confusions that may provide additional robustness, particularly in high word-error rate scenarios. We make use of log-linear models, which have been used for tagging confusion networks (Kurata et al., 2012), and which allow very flexible feature vector definition and discriminative optimization.

The model allows us to experiment with three types of inputs as illustrated in the Figure 3 – the manual transcript, the 1-best ASR transcript, and the word confusion network. To create supervised training data, we force-align ASR transcripts to manual transcripts and transfer manually annotated story element tags from the reference transcripts to word positions in the confusion network or 1-best ASR output using the word-level time marks. Our unsupervised training scenario instead derives story element tags from a baseline word-alignment based model.

Figure 3: Feature vectors at each word position includes lexical variants and acoustic confusions.



	Markov order 0 (MaxEnt)	Markov order 1 (CRF)
Context independent (CI)	y_i	$y_{i-1}y_i$
	$y_i x_i$	$y_{i-1}y_i x_i$
Context dependent (CD)	$y_i x_{i-1}$	$y_{i-1}y_i x_{i-1}$
	$y_i x_{i+1}$	$y_{i-1}y_i x_{i+1}$

Table 1: Feature templates either using or not using neighboring tag y_{i-1} (MaxEnt vs. CRF); and for using or not using neighboring words x_{i-1}, x_{i+1} (CI vs. CD).

4.1 Features

Given a sequence of word positions $x = x_1 \dots x_n$, the tagger assigns a sequence of labels $y = y_1 \dots y_n$ from a tag lexicon. For each word x_i in the sequence, we can define features in the log-linear model based on word and tag identities. Table 1 presents several sets of features, defined over words and tags at various positions relative to the current word x_i and tag y_i and compound features are denoted as concatenated symbols.

Features that rely only on the current tag y_i are used in a Markov order 0 model, i.e., one for which each tag is labeled independently. A maximum entropy classifier (see Section 4.2) is used with these feature sets. Features that include prior tags encode dependencies between adjacent tags, and are used within conditional random fields models (see Section 4.3). To examine the utility of surrounding words x_{i-1} and x_{i+1}, we distinguish between models trained with context independent features (just x_i) and context dependent features. Note that models including context dependent feature sets also include the context independent features, and Markov order 1 models also include Markov order 0 features.

Two other details about our use of the feature templates are worth noting. First, when tagging confusion networks, each word in the network at position i results in a feature instance. Thus, if there are five confusable words at position i, then there will be five different x_i values being used to instantiate the features in Table 1. Second, following Kurata et al. (2012), we multiply the feature counts for the context dependent features by a weight to control their influence on the model. In this paper, the scaling weight of the context-dependent features was 0.3.

We investigate two different tagsets for this task, as presented in Table 2. The simpler tagset (IO) simply identifies words that are in a story element; the

Tagging	*anna*	*rent was due*
IO-tags	I1	I19 I19 I19
BIO-tags	B1	B19 I19 I19

Table 2: Two possible tagsets for labeling.

larger tagset (BIO) differentiates among positions in a story element chunk. The latter tagset is only of utility for models with Markov order greater than zero, and hence are only used with CRF models.

4.2 MaxEnt-based multiclass classifier

Our baseline model is a Maximum Entropy (MaxEnt) classifier where each position i from the retelling x gets assigned one of the IO output tags y_i corresponding to the set of 25 story elements and a null ('O') symbol. The output tag is modeled as the conditional probability $p(y_i \mid x_i)$ given the word x_i at position i in the retelling.

$$p(y_i \mid x_i) = \frac{exp\left(\sum_{k=1}^{d} \lambda_k \phi_k(x_i, y_i)\right)}{Z(x_i)}$$

where $Z(x_i)$ is a normalization factor. The feature functions $\phi(x_i, y_i)$ are the Markov order 0 features as defined as in the previous section. The parameters $\lambda \in \Re^d$ are estimated by optimizing the above conditional probability, with L2 regularization. We use the MALLET Toolkit (McCallum, 2002) with default regularization parameters.

4.3 CRF-based sequence labeling model

The MaxEnt models assign a tag to each position from the input retelling independently. However, there are a few reasons why reframing the task as a sequence modeling problem may improve tagging performance. First, some of the story elements are multiword sequences, such as *she had been held up* or *on State Street*. Second, even if a retelling orders recalled elements differently than the original narrative, there is a tendency for story elements to occur in certain orders.

The parameters of the CRF model, $\lambda \in \Re^d$ are estimated by optimizing the following conditional probability:

$$P(y \mid x) = \frac{\exp\left(\sum_{k=1}^{d} \lambda_k \phi_k(x, y)\right)}{Z(x)}$$

where $\Phi(x, y)$ aggregates features across the entire sequence, and $Z(x)$ is a global normalization constant over the sequence, rather than local for a particular position as with MaxEnt. Features for the CRF model are Markov order 1 features, and as with the MaxEnt training, we use default (L2) regularization parameters within the MALLET toolkit.

5 Combining tagging and alignment

This paper contrasts a discriminatively trained tagging approach with an unsupervised alignment-based approach, but there are several ways in which the two approaches can be combined. First, the alignment model is unsupervised and can provide its output as training data to the tagging approach, resulting in an unsupervised discriminative model. Second, the alignment model can provide features to the log-linear tagging model in the supervised condition. We explore both methods of combination here.

5.1 Unsupervised discriminative tagger

The tagging task based on log-linear models provides an appropriate framework to easily incorporate diverse features and discriminatively estimate the parameters of the model. However, this approach requires supervised tagged training data, in this case manual labels indicating the correspondence of phrases in the retellings with story elements in the original narrative. These manual annotations are used to derive sequences of story element tags labeling the words of the retelling. Manually labeling the retellings is costly, and the scoring (thus labeling) scheme is very specific to the test being analyzed. To avoid manual labeling and provide a general framework that can easily be adopted in any retelling based assessment task, we experiment here with an unsupervised discriminative approach.

In this unsupervised approach, the labeled training data required by the log-linear model is provided by the automatic word alignments trained without supervision. The resulting tag sequences replace the manual tag sequences used in the standard supervised approach.

5.2 Word-alignment derived features

When training discriminative models it is a common practice to incorporate into the feature space the output from a generative model, since it is a good esti-

mator. Here we augment the feature space of the log-linear models with the tags generated by the automatic word alignments. In addition to the features defined in Section 4.1, we include new features that match predicted labels z_i from the word-alignment model with possible labels in the tagger y_i. Our features include the current tagger label with (1) the current predicted word-alignment label; (2) the previous predicted label; and (3) the next predicted label. Thus, the new features were $y_i z_i$, $y_i z_{i-1}$ and $y_i z_{i+1}$.

6 Experimental evaluations

Corpus: Our models were trained on immediate and delayed retellings from 144 subjects with a mean age of 85.4, of whom 36 were clinically diagnosed with MCI (training set). We evaluated our models on a set of retellings from 70 non-overlapping subjects with a mean age of 88.5, half of whom had received a diagnosis of MCI (test set). In contrast to the unsupervised word-alignment based method, the method outlined here required manual story element labels of the retellings. The training and test sets from this paper are therefore different from the sets used in previous work (Lehr et al., 2012; Prud'hommeaux and Roark, 2012; Prud'hommeaux and Roark, 2011), and the results are not directly comparable.

The recordings were sometimes made in an informal setting, such as the subject's home or a senior center. For this reason, there are often extraneous noises in the recordings such as music, footsteps, and clocks striking the hour. Although this presents a challenge for ASR, part of the goal of our work is to demonstrate the robustness of our methods to noisy audio.

6.1 Automatic transcription

The baseline ASR system used in the current work is a Broadcast News system which is modeled after Kingsbury et al. (2011). Briefly, the acoustics of speech are modeled by 4000 clustered allophone states defined over a pentaphone context, where states are represented by Gaussian mixture models with a total of 150K mixture components. The observation vectors consist of PLP features, stacked from 10 neighboring frames and projected to a 50-dimension space using linear discriminant analysis (LDA). The acoustic models were trained on 430 hours of transcribed speech from Broadcast News corpus (LDC97S44, LDC98S71). The language model is defined over an 84K vocabulary and consists of about 1.8M, 1M and and 331K bigrams, trigrams and 4-grams, estimated from standard Broadcast news corpus. The decoding is performed in several stages using successively refined acoustic models – a context-dependent model, a vocal-tract normalized model, a speaker-adapted maximum likelihood linear regression (MLLR) model, and finally a discriminatively trained model with the boosted MMI criteria (Povey et al., 2008). The system gives a word error rate of 13.1% on the 2004 Rich Transcription benchmark by NIST (Fiscus et al., 2007), which is comparable to state-of-the-art for equivalent amounts of acoustic training data. On the WLM corpus, the recognition word error rate was significantly higher at 47.2% due to a mismatch in domain and the skewed demographics (age) of the speakers.

System	1-best (%)	oracle WCN(%)	oracle lat(%)
Baseline	47.2	39.7	27.7
AM adaptation	38.2	35.5	21.2
LM adaptation	28.3	30.7	19.9
AM+LM adaptation	25.6	26.5	16.5

Table 3: Improvement in ASR word error-rate by adapting the Broadcast News models to the domain of narrative retelling.

We improved the performance of the above Broadcast News models by adapting to the domain of the WLM retellings. The acoustic models were adapted using standard MLLR, where linear transforms were estimated in an unsupervised manner to maximize the likelihood over the transcripts of the retellings. The transcripts were generated from the baseline system after the final stage of decoding with the discriminative model. The language models were adapted by interpolating the in-domain model (weight=0.7) with the out-of-domain model. The gains from these adaptations are reported in the Table 3. As expected, we find substantial gains from both acoustic model (AM) and language model (LM) adaptation. Furthermore, we find benefit in employing them simultaneously. We also include the oracle word error rate (WER) of the WCNs and lattices for each ASR configuration.

One thing to note is that the oracle WER of the WCNs is worse than the 1-best WER when adapting the language models. We speculate that this is due to bias introduced by the language model adapted to the story retellings, resulting in word candidates in the bins that are not truly acoustically confusable candidates. This is one potential reason for the lack of utility of WCNs in low WER conditions.

6.2 Evaluating retelling scoring

We analyzed the performance of the retelling scoring methods under five different input conditions for producing transcripts: (1) the out-of-domain Broadcast News recognizer with no adaptation; (2) domain adapted acoustic model; (3) domain adapted language model; (4) domain adapted acoustic and language models; and (5) manual (reference) transcripts. Each story element is automatically labeled by the systems as either having been recalled or not, and this is compared with manual scores to derive an F-score accuracy, by calculating precision and recall of recalled story elements. Derived word alignments or tag sequences are converted to binary story element indicators by simply setting the element to 1 if any open-class word is tagged for (or aligned to) that story element.

6.2.1 Word alignment based scoring

We evaluate the word alignment approach only on 1-best ASR transcripts and manual transcripts, not WCNs. The first row of Table 4 reports the story element F-scores for a range of ASR adaptation scenarios. The performance of the model improves significantly as the WER reduces with adaptation. With the fully adapted ASR the F-score improves more than 13%, and it is only 3.4% worse than with the manual transcripts. The alignments produced in each of these scenarios are used as training data in the unsupervised condition evaluated below.

6.2.2 Log-linear based automated scoring

Context-independent features Table 4 summarizes the performance of the log-linear models using context independent features (CI) in supervised (section 4), unsupervised (section 5.1) and hybrid (section 5.2) training scenarios for different inputs (reference transcript, ASR 1-best, and word confusion network ASR output) and four different ASR configurations.

The results show a few clear trends. Both in the supervised and unsupervised training scenarios the CRF model provides substantial improvements over the MaxEnt classifier. The F-scores obtained in the unsupervised training scenario are slightly worse than with supervision, though they are comparable to supervised results and an improvement over just using the word alignment approach, particularly in high WER scenarios. The hybrid training scenario – supervised learning with word alignment derived features – leads to reduced differences between MaxEnt and CRF training compared to the other two training scenarios. In fact, in high WER scenarios, the MaxEnt slightly outperforms the CRF.

As expected the best performance is obtained with manual transcripts and the worst with 1-best transcripts generated by the out-of-domain ASR with relatively high word error rate. For this ASR configuration, using WCNs provide some gain, though the gain is insignificant for the hybrid approach. In the hybrid approach, the output labels of the word alignment are already good indicators of the output tag and incorporating the confusable words from the

Table 4: Story element F-score achieved by baseline word-alignment model and log-linear models (MaxEnt and CRF) using **context independent** features (CI) under 3 different scenarios, with 3 different inputs (1-best ASR, word confusion network, and manual transcripts) and different ASR models (baseline out-of-domain, AM adapted, LM adapted and AM+LM adapted).

Training	Transcripts:	1-best				WCN				manual
Scenario	ASR:	baseline	AM	LM	AM+LM	baseline	AM	LM	AM+LM	N/A
Baseline word-alignment:		71.9	77.3	84.3	85.4	N/A				88.8
Supervised	MaxEnt-CI	76.0	81.7	84.6	85.6	78.9	83.4	84.0	84.7	86.4
	CRF-CI	80.3	87.3	89.7	91.4	83.7	88.8	88.2	90.8	94.4
Unsupervised	MaxEnt-CI	72.1	79.3	82.7	84.2	77.5	81.2	83.4	83.2	84.8
	CRF-CI	79.4	85.4	86.8	88.0	81.2	85.8	86.2	87.2	90.5
Hybrid	MaxEnt-CI	88.1	89.4	89.2	89.6	87.6	89.2	88.8	89.5	91.8
	CRF-CI	87.0	90.9	91.5	92.1	87.4	91.5	90.1	92.4	94.6

Training Scenario	Transcripts: ASR:	1-best baseline	1-best AM	1-best LM	1-best AM+LM	WCN baseline	WCN AM	WCN LM	WCN AM+LM	manual N/A
Supervised	MaxEnt-CD	80.1	87.3	90.0	91.1	83.5	88.6	88.2	90.3	93.3
	CRF-CD-IO	80.6	88.0	89.9	91.2	84.2	89.6	88.8	90.5	94.7
	CRF-CD-BIO	81.1	87.9	90.6	91.7	84.5	89.5	88.8	90.8	94.7
Un-supervised	MaxEnt-CD	77.1	83.1	86.5	89.0	80.2	85.0	86.2	87.6	90.7
	CRF-CD-IO	79.1	85.3	87.1	88.3	81.0	85.9	86.4	87.5	90.3
	CRF-CD-BIO	79.1	85.6	87.2	88.4	81.3	85.9	86.2	87.3	90.6
Hybrid	MaxEnt-CD	88.4	90.2	90.7	91.6	88.6	90.5	90.4	91.4	93.5
	CRF-CD-IO	87.9	91.3	91.6	92.5	88.3	91.7	90.7	92.1	94.8
	CRF-BIO	87.8	91.9	91.8	93.0	88.7	92.0	90.7	92.3	94.7

Table 5: Story element F-score achieved by log-linear models (MaxEnt and CRF) when adding **context dependent** features (CD) and **BIO tags** for the CRF models, under 3 different scenarios, with 3 different inputs (1-best ASR, word confusion network, and manual transcripts) and different ASR models (baseline out-of-domain, AM adapted, LM adapted and AM+LM adapted).

WCN into the feature vector apparently mainly adds noise.

When the transcripts are generated with the adapted models, the word confidence score of the 1-best is higher and the WCN bins have fewer acoustically confusable words. Still, the WCN input is helpful in the AM-adapted ASR system. When the transcripts are generated with LM adapted models, the performance is better with 1-best than with WCNs. As mentioned earlier, adapting the language models may introduce a bias due to the relatively low LM perplexity for this domain. In the lowest WER scenarios, the best performing systems achieve over 90% F-score, within two percent of the performance achieved with manual transcripts.

Context-dependent features Exercising the flexibility of log-linear models, we investigated the impact of using context-dependent (CD) features instead of the CI features used in the previous experiments. Our CD features take into account the two immediately neighboring word positions. As mentioned earlier, following Kurata et al. (2012), the counts from the neighboring word positions were weighted ($\alpha = 0.3$) to avoid data sparsity. This reduces the sensitivity of the model to time alignment errors between the tag and feature vector sequences without increasing the dimensions. In Table 5, we report the F-scores for the different ASR configurations, inputs, and log-linear models with context dependent features, using the standard IO tagset as in Table 4.

Although there are some exceptions, adding context information from the input features improves the performance of the models. In particular, the MaxEnt models benefit from incorporating this extra information. The MaxEnt models improve their performance substantially for all three training scenarios, while the gains for the CRF models are more modest, especially for the unsupervised approach where the performance degrades or does not change much, since some context information is already captured by the Markov order 1 features.

BIO tagset As detailed in Section 4.1, story elements sometimes span multiple words, so for the CRF models we investigated two different schemes for tagging, following typical practice in named entity extraction (Ratinov and Roth, 2009) and syntactic chunking (Sha and Pereira, 2003). The BIO tagging scheme makes the distinction between the tokens from the story elements that are in the beginning from the ones that are not. The O tag is assigned to the tokens that do not belong to any of the story elements. The IO tagging uses a single tag for the tokens that fall in the same story element, which is the approach we have followed so far. In addition to presenting results using context dependent features, Table 5 presents results with the BIO tagset.

For the supervised and hybrid approaches, the BIO tagging provides insignificant but consistent gains for most of the scenarios. The unsupervised approach provides mixed results. This may be due to the way in which the word alignment model scores the retellings. It tags only those words from the retelling that are aligned with a content word in the source narrative, which may result in the loss of the

Training Scenario	Transcripts:	1-best				WCN				manual
	ASR:	baseline	AM	LM	AM+LM	baseline	AM	LM	AM+LM	N/A
Baseline word-alignment:		0.65	0.67	0.74	0.76	N/A				0.79
Supervised	MaxEnt-CD	0.65	0.73	0.76	0.77	0.70	0.73	0.77	0.77	0.81
	CRF-CD-BIO	0.69	0.76	0.77	0.76	0.73	0.76	0.77	0.78	0.82
Un-supervised	MaxEnt-CD	0.65	0.72	0.75	0.76	0.70	0.75	0.75	0.76	0.80
	CRF-CD-BIO	0.74	0.75	0.78	0.78	0.71	0.74	0.77	0.76	0.81
Hybrid	MaxEnt-CD	0.72	0.76	0.77	0.78	0.74	0.76	0.77	0.77	0.82
	CRF-CD-BIO	0.72	0.76	0.78	0.78	0.76	0.77	0.78	0.79	0.81

Table 6: Classification performance (AUC) for the baseline word-alignment model and the best performing log-linear models of both types (MaxEnt and CRF) under 3 different scenarios with 3 types of input and 4 types of ASR models.

structure of some multiwords story elements that we are trying to capture with the BIO scheme.

6.3 Evaluating MCI classification

Each of the individuals producing retellings in our corpus underwent a battery of neuropsychological tests, and were assigned a Clinical Dementia Rating (CDR) (Morris, 1993), which is a composite score derived from measures of cognitive function in six domains, including memory. Importantly, it is assigned independently of the Wechsler Logical Memory test we are analyzing in this paper, which allows us to evaluate the utility of our WLM analyses in an unbiased manner. MCI is defined as a CDR of 0.5 (Ritchie and Touchon, 2000), and subjects in this study have either a CDR of 0 (no impairment) or 0.5 (MCI).

In previous work, we found that the features extracted from the retellings are useful in distinguishing subjects with MCI from neurotypical age-matched controls (Lehr et al., 2012; Prud'hommeaux and Roark, 2012; Prud'hommeaux and Roark, 2011). From each retellings, we extract Boolean features for each story element, for a total of 50 features for classification. Each feature indicates whether the retelling contained that story element.

In this paper, we carry out similar classification experiments to investigate the impact of using log-linear models on the extraction of features for classification. We build a support vector machine (SVM) using the LibSVM (Chang and Lin, 2011) extension to the WEKA data mining Java API (Hall et al., 2009). This allows recollection of different elements to be weighted differently. This is unlike the manual scoring of WLM based on clinical guidelines where all elements are weighted equally irrespective of the difficulty. The SVM was trained on manually extracted story element feature vectors. We compared the performance of the MCI classification for three types of input and four ASR configurations under the supervised, unsupervised, and hybrid scenarios. For each scenario we chose the best scoring system from among the automated systems reported in Tables 4 and 5. Classification results, evaluated as area under the curve (AUC), are reported in Table 6, both for the log-linear trained tagging models and for the baseline word-alignment based method. For reference (not shown in the table), the SVM classifier performed at 0.83 when features values are manually populated.

The results show that the AUC improves steadily as the quality of the transcription is improved, going from the baseline system to the adapted models. This is consistent with the improvements seen in the F-score for detecting story elements. The different approaches for detecting the story elements from the transcriptions did not ultimately show significant differences in MCI classification results. Overall, the best classification values are given by the hybrid approach, which performs slightly better than the other two approaches. The best AUC in the hybrid scenario (0.79, very close to the AUC=0.81 achieved with manual transcripts) is obtained with a CRF trained with WCNs from the fully adapted ASR model and with context dependent features and BIO tags.

Comparing WCN versus 1-best as inputs, using WCN as input improves classification performance when the 1-best transcripts are poor, as in the case of out-of-domain ASR. The adapted recognizer improves the performance of the 1-best significantly making it unnecessary to resort to WCN as inputs.

Comparing the MaxEnt model with CRF model

for extracting story elements, we see that the average F-scores for the MaxEnt models trained on CD features are nearly as good as and sometimes slightly better than those produced using the CRF models. The CRF extracted story elements, however, tend to yield classifiers that perform slightly better, especially in the unsupervised approach with 1-best inputs.

7 Summary and discussion

This paper examines the task of automatically scoring narrative retellings in terms of their fidelity to the original narrative content, using discriminatively trained log-linear tagging models. Fully automatic scoring must account for both lexical variation and acoustic confusion from ASR errors. Lexical variation – due to extensive paraphrasing on the part of the individuals retelling the narrative – can be modeled effectively using word-alignment models such as those employed in machine translation systems (Lehr et al., 2012; Prud'hommeaux and Roark, 2011). This paper focuses on an alternative approach, where both lexical variation and ASR confusions are modeled using log-linear models. In addition to very flexible feature definitions, the log-linear models bring the advantage of a discriminative model to the task. We see improvements in story element F-score using these models over unsupervised word-alignment models. Further, the feature definition flexibility allows us to incorporate the unsupervised word-alignment labels into these models, resulting either in fully unsupervised approaches that perform competitively with the supervised models or in hybrid (supervised) approaches that provide the best performing systems in this study.

Our tagging models are able to process word confusion networks as inputs and thus improve performance over using 1-best ASR transcripts in scenarios where the speech recognition error rate is high. These improvements carry through to the MCI classification task, making use of features computed from the automatic scoring of narrative retelling.

One advantage of the word-alignment model is that such approaches do not require manual annotation of the story elements, which is more labor intensive than typical manual transcription of speech. Thus, the word-alignment model can exploit large numbers of retellings in an unsupervised manner when trained on ASR transcripts of the retellings. Controlled experiments here with relatively limited training sets demonstrate that semi-supervised approaches on larger untranscribed sets are likely to be successful.

Finally, experiments with different amounts of ASR adaptation show that both acoustic and language model adaptations in this domain are effective, yielding scenarios that are competitive with manual transcription both for detecting story elements as well as for subsequent classification. With full model adaptation to the domain, the 1-best transcripts improved significantly, and their performance was found to be at par with WCNs.

In future work, we would like to investigate two questions left open by these results. First, word-alignment models can be extended to process ASR lattices or word confusion networks as part of the unsupervised alignment learning algorithm, and incorporated into our approach. Second, the contextual features can be refined (e.g., concatenated features instead of smoothed features) when large amounts of training data is available.

It is noteworthy to mention that the lexical variants and paraphrasing learned from the data using automated method may be useful in refining the clinical guidelines for scoring (e.g., allowing additional lexical variants and paraphrasings, or assigning unequal credits for different story elements to reflect the difficulty of recollecting them) or to create the guidelines for new languages or stories.

Acknowledgments

This research was supported in part by NIH awards 5K25AG033723-02 and P30 AG024978-05 and NSF awards 1027834, 0958585, 0905095, 0964102 and 0826654. 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. We thank Brian Kingsbury and IBM for the use of their ASR software tools; Jeffrey Kaye and Diane Howeison for their valuable input; and the clinicians at the Oregon Center for Aging and Technology for their care in collecting the data.

References

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.

Jonathan Fiscus, John Garofolo, Audrey Le, Alvin Martin, greg Sanders, Mark Przybocki, and David Pallett. 2007. 2004 spring nist rich transcription (rt-04s) evaluation data. `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2007S12`.

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).

Brian Kingsbury, Hagen Soltau, George Saon, Stephen M. Chu, Hong-Kwang Kuo, Lidia Mangu, Suman V. Ravuri, Nelson Morgan, and Adam Janin. 2011. The IBM 2009 GALE Arabic speech transcription system. In *Proceedings of ICASSP*, pages 4672–4675.

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

Gakuto Kurata, Nobuyasu Itoh, Masafumi Nishimura, Abhinav Sethy, and Bhuvana Ramabhadran. 2012. Leveraging word confusion networks for named entity modeling and detection from conversational telephone speech. *Speech Communication*, 54(3):491–502.

Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Interspeech*.

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, Los Angeles.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

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(11):1485–1490.

Ronald Petersen, 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.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah. 2008. Boosted mmi for model and feature space discriminative training. In *Proceedings of ICASSP*.

Emily Prud'hommeaux and Brian Roark. 2011. Alignment of spoken narratives for automated neuropsychological assessment. In *Proceedings of ASRU*.

Emily Prud'hommeaux 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)*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *EMNLP*.

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

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*.

Martha Storandt and Robert Hill. 1989. Very mild senile dementia of the Alzheimer's type: II Psychometric test performance. *Archives of Neurology*, 46:383–386.

Qing-Song Wang 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.

David Wechsler. 1997. *Wechsler Memory Scale - Third Edition*. The Psychological Corporation, San Antonio.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of HLT-NAACL*.

Xiaonan Zhang, Xiaonan Zhang, Jack Mostow, Jack Mostow, Nell Duke, Christina Trotochaud, Joseph Valeri, and Al Corbett. 2008. Mining free-form spoken responses to tutor prompts. In *Proceedings of the First International Conference on Educational Data Mining*, pages 234–241.