Linguistically Engineered Tools for Speech Recognition Error Analysis

Carol Van Ess-Dykema[1] *Klaus Ries*[2]
cjvanes@afterlife.ncsc.mil ries@cs.cmu.edu

[1]U.S. Department of Defense 9800 Savage Rd., Ft. Meade, Maryland 20755
[2]Interactive System Labs, Carnegie Mellon University, USA and University of Karlsruhe, Germany

ABSTRACT

In order to improve Large Vocabulary Continuous Speech Recognition (LVCSR) systems, it is essential to discover exactly how our current systems are underperforming. The major intellectual tool for solving this problem is error analysis: careful investigation of just which factors are contributing to errors in the recognizers. This paper presents our observations of the effects that discourse (i.e., dialog) modeling has on LVCSR system performance. As our title indicates, we emphasize the recognition error analysis methodology we developed and what it showed us as opposed to emphasizing development of the discourse model itself. In the first analysis of our output data, we focussed on errors that could be eliminated by Dialog Act discourse tagging [JSB97] using Dialog Act-specific language models. In a second analysis, we manipulated the parameterization of the Dialog Act-specific language models, enabling us to acquire evidence of the constraints these models introduced. The word error rate did not significantly decrease since the error rate in the largest category of Dialog Acts, namely Statements, did not significantly decrease. We did, however, observe significant error reduction in the less frequently occurring Dialog Acts and we report on the characteristic of the error corrections. We discovered that discourse models can introduce simple syntactic constraints and that they are most sensitive to parts of speech.

1. Introduction

In speech recognition research there are two traditions for identifying error sources. One approach we call the engineering or statistical approach. This approach focuses the attention on a single metric, optimally on a single number such as word accuracy or perplexity, which can be calculated automatically. On the other end of the spectrum, the comparison of machine with human transcripts is used to determine properties of errors the LVCSR system is producing. We call this the "language expert's approach".

Neither approach is ideal. The engineering or statistical approach is hard to interpret and laborious to implement. Additionally, it often only verifies a hypothesis and rarely generates new hypotheses. The language expert's approach takes a long time to carry out and reading the transcripts and assigning error sources can be tedious and confusing. On top of this, the evidence generated is often small and may be irrelevant.

Our first experiments were part of the Johns Hopkins University Center for Language and Speech Processing (CLSP) LVCSR Summer Workshop'97 as members of the team working on discourse modeling for the Switchboard (SWBD) corpus [JSB97]. The corpus was already segmented into Dialog Act units and the task was to automatically assign correct tags. The question we wanted to answer was how to evaluate the effect of discourse modeling on LVCSR word accuracy. The work on discourse classification was continued in project Clarity at Carnegie Mellon University. The setup in Clarity [FLL+98] is slightly different from that in the Johns Hopkins Workshop. The data has not been presegmented into Dialog Acts and the Dialog Act classifier has to perform both the classification and segmentation tasks. Additionally, Clarity is using the CallHome Spanish database and the number of Dialog Act-tagged dialogs is an order of magnitude smaller than in the Switchboard database.

Our first data were "cheating" runs carried out by our team at the Johns Hopkins LVCSR summer workshop, where the language model used for decoding was determined based on the manually assigned Dialog Acts. Our main goal was to determine how to use the Dialog Act classification for improving word accuracy. We found that in addition to word accuracy and unsupported transcript reading, we needed a "linguistically engineered" viewpoint to best analyze the word recognition errors (Figure 1). We present our analysis methodology and results in detail in Sec. 2.

Later, in Clarity, we used a more classical technique known in AI research. We modified the "input representation" of the classifier and assessed the effect these inputs have on the classifiers (Sec. 3).

Neither the work at the LVCSR workshop nor in Project Clarity is unique in its use of Dialog Act detection technology. It has been proposed and used by various other groups such as [NM94, WKNN97, REKK96, TKI+97]. Both the English SWBD and the CallHome Spanish databases, however, differ from the more task-oriented dialogs that have been used up until now in the speech and discourse community. In the SWBD database, the telephone speakers are unknown to each other and have been instructed to talk with each other about a specific topic. CallHome speakers call their home country and chat with their family members. We hope that the effects of our discourse models and the error analysis methodology we developed will help us compare and analyze the various classifiers developed to-date and that they can be extended to enable a comparison between different discourse styles (e.g.,

task-based vs. spontaneous speech).

In our Conclusion and Future Research section, (Section 4) we include additional areas where we think further linguistic analysis can be helpful and we suggest methods for achieving them. Some of these tasks are being addressed within project Clarity and the different data collections mentioned are ongoing or are already complete.

2. Error Analysis Tool

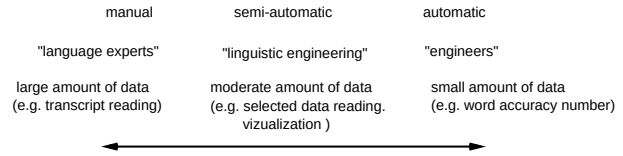

Figure 1: Viewing error analysis as a continuum: In addition to word accuracy percentages and to unsupported transcript reading, an intermediate view of the data is important.

```
REF:  DO YOU HAVE   SMOG LIKE   THEY DO     in  california
HYP:  ** *** YOU'RE NOT  SMOKE  I    COULD  in  california
REF:  do you have SMOG LIKE THEY   DO      in california
HYP:  do you have **** **** SMOKE  ACTIVE  in california

REF: -DO YOU REALLY think CARS CONTRIBUTE [..]
HYR: *** AND THEY   think **** ARE        [..]
REF: -do you REALLY think CARS CONTRIBUTE [..]
HYP: do  you ****** think **** OUR        [..]

REF: DO YOU
HYP: ** OKAY
REF: do you
HYP: do you
```

Figure 2: Looking at selected data: When we decided to look for the improvements the Dialog Act-specific language model could achieve and selected Questions, the Question-initial word *do* was very prominent. The figure shows four utterances: The top reference/hypothesis pair uses the baseline language model, while the bottom pair uses a Question-specific language model.

The task of the error analysis tool was to extract information from "cheating" runs of our discourse models. The cheating runs were produced at the Johns Hopkins LVCSR summer workshop and used the manually assigned Dialog Act information to condition the language model [JBC+ 97].

Our tool allows us to select, combine, group and display information from different LVCSR outputs, manual transcriptions, and other sources at the Dialog Act level. The tool also incorporates alignment information from the LVCSR output and the reference transcripts. The most important grouping feature is the manual Dialog Act classification of the utterance. The most useful function is the display aligning the reference transcript with the LVCSR output using a standard language model and with the output using a language model conditioned on the (manually determined) Dialog Act (Figure 2). This mode allowed us to browse large amounts of data and it showed us e.g. that the initial portion of the utterance is often corrected.

To verify this trend in the data we compared the error rate over the whole turn (all) with the error rate in the first three words (initial) in the baseline system. We contrasted the baseline language model with the relative improvement we gained from using Dialog Act-specific language models (cheating). We found that the Dialog Act-specific model frequently corrects errors at the beginning of the utterance (Table 1). We assume that this is due to the simple syntactic constraints frequently found in surface realizations of the Dialog Acts, e.g. the Question-initial word do. For the spontaneously spoken SWBD dialogs (and the CallHome dialogs), however, the major Dialog Act category does not exhibit much improvement and the overall effect on word accuracy is therefore small. Most of the remaining Dialog Acts do show significant improvements, however.

Dialog Act	Word Accuracy Baseline all	Word Accuracy Baseline initial	Improvement Cheating all	Improvement Cheating initial
Statement	58.1	58.9	0.71	1.20
Backchannel	78.2	78.2	7.72	7.53
Opinion	59.6	57.0	0.47	1.07
Abandoned	52.3	50.8	7.79	11.64
Agree/Accept	81.1	83.6	-1.22	0.69
SWBD	58.6	57.4	-1.49	3.52

Table 1: Turn-initial improvements from Dialog Act knowledge: We compare the error rate over the whole turn (all) with the error rate in the first three words (initial) in the baseline system and contrast that with the relative improvement we gain from using Dialog Act-specific language models (cheating). The overall trend is that the Dialog Act-specific model corrects errors at the beginning of the utterance (originally reported in [JBC+ 97]).

Combining Our Experiment Results Into a Single Matrix Our discourse language model had two goals: 1) to automatically detect the Dialog Acts of utterances and 2) to constrain the language model to the Dialog Act-specific model. We asked the following questions:

- Does the LVCSR system detect the words that discriminate between Dialog Acts?
- Which Dialog Acts are discriminated?
- Which words frequently discriminate and do they correlate with the Dialog Act type?
- Are there Dialog Act-specific frequently occurring words that are often wrongly recognized?

If we were working with higher order n-gram models, a manual analysis of the Dialog Act detection model would not be feasible. We therefore primarily used unigrams enriched with approximately 190 multi-words like YOU_KNOW that have been used in LVCSR systems in the recent past. We used the frequency of words, word salience [GLG91], and word error rate per Dialog Act as measures and combined them into a single table (see Table 2).

Word ranked by frequency					Words ranked by sal.
Word	Frequency	Salience	Salience Rank	Word Error	
Statement-Non-opinion					
and	18116	0.31816	2	64%	THE
the	17570	0.34334	1	27%	AND
I	14600	0.22314	8	50%	UH
uh	14250	0.30139	3	0%	YEAH
that	13538	0.29036	5	65%	THAT
Acknowledge (Backchannel)					
uh_huh	14446	2.17960	1	66%	UH_HUH
yeah	13776	2.12916	2	33%	YEAH
right	3583	0.73160	3	85%	RIGHT
oh	2543	0.60609	4	12%	OH
okay	770	0.34320	8	65%	UH
Statement-Opinion; Explicit Performative					
the	8088	0.37340	1	65%	THE
that	7232	0.32480	3	56%	AND
and	5870	0.32701	2	35%	THAT
to	5399	0.26613	5	93%	UH
uh	5234	0.30052	4	53%	TO
Abandoned/Turn-Exit; Uninterpretable					
uh	2202	0.78549	1	32%	UH
so	2120	0.71340	2	52%	SO
but	1635	0.62159	4	50%	AND
and	1490	0.67716	3	27%	BUT
I	1120	0.57710	5	0%	I

Table 2: Error Matrix: The error matrix shows that the salient words are often Frequently occurring words and vice versa. Some of the salient words are really short and are often misrecognized. The matrix combines many different analyses into one scheme – it is a compact presentation of the results of our experiments.

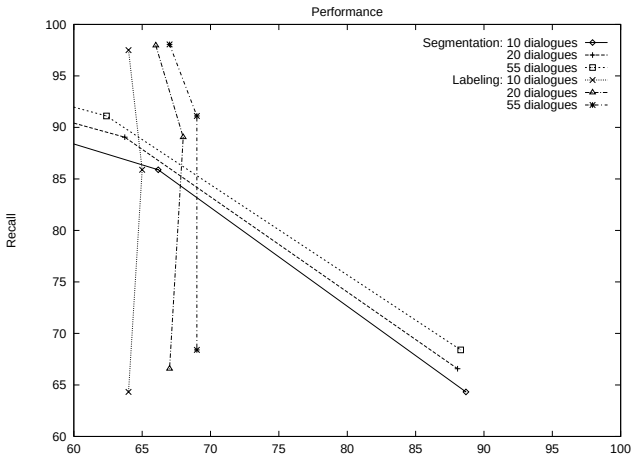

Figure 3: Varying the size of the training corpus: A good Dialog Act segmenter and classifier can be trained from as few as 30 dialogs. The figure shows the precision and recall of the segmenter (the vertical lines in the left part of the figure) and for each recall of the segmenter, the Dialog Act classification precision measured on the word level (the lines going from the upper left to the lower right corner). See also Figure 4.

3. Classifier Experiments

In the second set of experiments a Dialog Act classifier was trained for CallHome Spanish dialogs. We found that the language models fairly accurately model the length distribution of the corresponding Dialog Act utterances (reported in [FLL+98]). We therefore concluded that any integration of prosodic models with language models has to take this into account, especially since the length of the utterance is the strongest prosodic cue available and most other cues are dependent on this one [SBC+98].

This set of experiments allows us to report what these classifiers actually learn. First, it became very obvious that the classifiers can already achieve most of their performance from 30 dialogs. Significantly fewer dialogs still have a strong effect on the accuracy, and significantly more do not improve the accuracy (Figure 3). This could be indicative of the type of rule learned: The characteristic has to be really frequent in the database, there are few characteristics overall, and they might be fairly simple.

We used the parts of speech (POS) of the words in the input sentence as the input to the classifier. However, along the lines of [SS96, GZA97] we did not map the most frequent words to their parts of speech but used the word/POS pair as an entry for the language model vocabulary alongside with the standard parts of speech [1]. Notwithstanding our small number of dialogs, we achieved good results; only a few words did not map onto their parts of speech (Figure 4). In another experiment, we tried to determine which words would not map to the raw POS tag by salience [GLG91]. If there were certain keywords which by their presence alone would determine the Dialog Act, this selection mechanism should deliver better classification results. However, we discovered that the frequency-based selection consistently outperformed the salience-based selection by a small margin.

[1] We used a modified version of [Bri93] and we acknowledge Klaus Zechner for building the tagger

4. Conclusion and Future Research

This paper demonstrates how we have exploited computational tools to advantageously use linguistic techniques in conjunction with statistical dialog processing and LVCSR technology. We presented a technique that allowed us to perform an effective analysis of LVCSR output using discourse categories (i.e., Dialog Acts). Our error analyses combined manually annotated discourse information with the alignment information from the speech recognizer. This technique can be extended by performing more linguistic annotation of the utterances and further partitioning of the data. We found that conditioning the language model on the Dialog Act typically yields an improvement for most Dialog Acts and that the improvements tend to be turn-initial. Our classifier experiments seem to indicate that the constraints are simple syntactic constraints, since the classification of Dialog Acts can be learned from parts of speech using small databases.

Another analysis we plan to carry out is to compare the CallHome Spanish, CallHome English, and SWBD English dialog corpora. We have performed Dialog Act annotations for the CallHome English corpus and are thus able to compare the difference in corpus style (CallHome English vs. SWBD English) with the difference in language (CallHome Spanish vs. CallHome English). We hope that this kind of study will give us more insight into everyday discourse and its different dimensions. Another dimension open to evaluation is a comparison with more task-oriented styles such as [TKI+97, WKNN97, REKK96] and others.

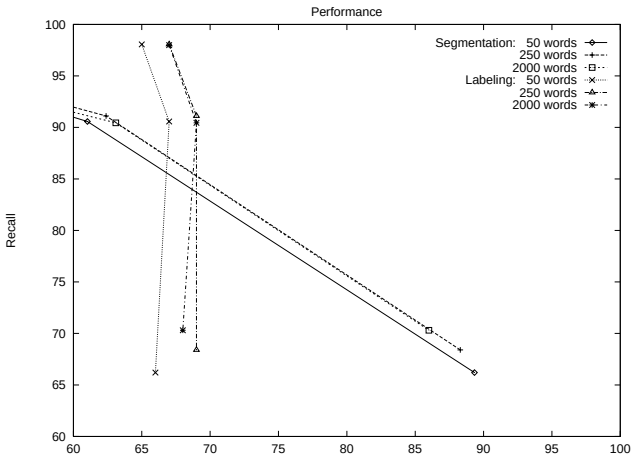

Figure 4: Varying the POS model: The performance of the models slowly improves as long as only a small number of words are mapped onto their parts of speech. The improvements level off fairly quickly but without any over-training effects. See also Figure 3.

Another future challenge is to include more discrimination capabilities in the major Discourse Act category, the Statements. [MZM98] focussed on subsegments of Statements and this technique has already shown significant word accuracy improvements. During the LVCSR summer workshop an initial study [JBC+97] showed that one can find different types of Statements according to their discourse context. We are actively pursuing this sub-categorization of Statements further within project Clarity [LTGR+98].

5. Acknowledgments

We would like to acknowledge the contributions to our work from our Johns Hopkins LVCSR 1997 Summer Workshop discourse team: Rebecca Bates, Noah Coccaro, Daniel Jurafsky (team leader), Rachel Martin, Marie Meteer, Elizabeth Shriberg, Andreas Stolcke and Paul Taylor. We would also like to thank our fellow Project Clarity members at Carnegie Mellon University: Michael Bett, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Alex Waibel and Klaus Zechner, as well as Ann Thyme-Gobbel and Sandra E. Hutchins from Natural Speech Technologies. Project Clarity is conducted under a contract by the U.S. Department of Defense.

We also thank the Discourse Act taggers at the University of Colorado, Boulder, and at Carnegie Mellon.

Lastly, we would like to acknowledge Susann LuperFoy from the MITRE Corp, Tom Crystal from IDA, and Cal Olano and Sara Shelton from the U.S. Department of Defense for their contributions and support.

6. REFERENCES

Bri93. E. Brill. *A Corpus-Based Approach to Language Learning*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, 1993.

FLL+98. Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. Clarity: Automatic discourse and dialogue analysis for a speech and natural language processing system. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998.

GLG91. A.L. Gorin, S. E. Levibson, and A. N. Gertner. Adaptive acquisition of spoken language. In *ICASSP*, pages 805–809, 1991.

GZA97. Marsal Gavalda, Klaus Zechner, and Gregory Aist. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Fifth Conference on Applied Natural Language Processing*, Washington,DC, 1997.

JBC+97. Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. SWBD Discourse Language Modeling Project, Final Report. Technical report, Johns Hopkins LVCSR Workshop-97, 1997.

JSB97. Dan Jurafsky, Liz Shriberg, and D. Biasca. Switchboard-damsl labeling project coder's manual. Technical report, Institute of Cognitive Science, University of Colorado, Boulder, USA, 1997.

LTGR+98. Lori Levin, Ann Thyme-Gobbel, Klaus Ries, Alon Lavie, and Monika Woszczyna. A discourse coding scheme for conversational spanish. In *ICSLP*, 1998.

MZM98. Kristine W. Ma, George Zavaliagkos, and Marie Meteer. Sub-sentence discourse models for conversational speech recognition. In *ICASSP*, 1998.

NM94. M. Nagata and T. Morimoto. First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203, 1994.

REKK96. N. Reithinger, R. Engel, M. Kipp, and M. Klesen. Predicting dialogue acts for a speech-to-speech translation system. In *ICSLP*, 1996.

SBC+98. Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, forthcoming, 1998.

SS96. A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In *Proceedings of the ICSLP*, Philadelphia, USA, 1996.

TKI+97. Paul Taylor, Simon King, Stephen Isard, Helen Wright, and Jacqueline Kowtko. Using intonation to constrain language models in speech recognition. In *EUROSPEECH*, Rhodes, Greece, 1997.

WKNN97. V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated dialog act segmentation and classification using prosodic features and language models. In *Eurospeech*, pages 207–210, 1997.