

# LEVERAGING STRUCTURAL MDE IN LANGUAGE PROCESSING

*M. Ostendorf, J. Kahn, D. Wong, D. Hillard and W. McNeill*

Signal, Speech and Language Interpretation Lab, University of Washington, Seattle, WA  
{mo, jgk, darby, hillard, billmcn}@ssli.ee.washington.edu

## ABSTRACT

This paper looks at the interaction between structural metadata and language structure, i.e. syntax, in conversational speech. We present results in the use of automatically detected boundary events to improve parsing performance, and argue that the results suggest promise for using these events in language modeling. We describe strategies for incorporating metadata-informed language modeling and present results in weakly supervised learning of prosodic constituent structure that are key to the approach.

## 1. INTRODUCTION

The lack of explicit sentence boundaries and presence of disfluencies pose difficulties for automatically parsing conversational speech, though humans typically do not find it particularly difficult. Humans make use of acoustic cues that are not expressed in a simple orthographic word transcription, cues that can be thought of as analogous to punctuation in written text, which we refer to as structural metadata. Since text processing systems productively make use of punctuation, it makes sense that language processing systems operating from speech would benefit from annotation of structural metadata in the word stream. This paper explores this question for two problems: parsing and language modeling.

In particular, we focus on the subclass of metadata that can be considered boundary events, i.e. a phenomenon that marks the edge of some constituent. Examples of such events might be boundaries of intonational phrases (prosodic constituents), sentences, self-interruption points, speaker change points and discourse segment boundaries. There is much work in linguistics showing that boundary events are important for human segmentation and disambiguation of language, so these cues are good candidates for initial attempts to integrate metadata into language processing systems. Further, the acoustic cues for boundary events (pauses, duration lengthening and boundary tones) are mostly local to the boundary rather than extending over the whole constituent, so they are well suited to current left-to-right speech recognition models.

In this study, we look at self-interruption points (IPs) and boundaries of sentence-like units (SUs) in conversational telephone speech, because of the availability of annotated data [1] and automatic detection systems [2, 3]. An SU roughly corresponds to a sentence, except that SUs are for the most part defined as units that include only one independent main clause, and they may sometimes be incomplete as when a speaker is interrupted and does not complete a sentence. There are four types of SUs annotated in our corpus (statement, question, backchannel, and incomplete), but much of our work uses only a binary SU vs. no SU distinction. An IP marks the end of a sequence of words that the speaker decides is in error. It may be followed by an explicit edit term or filled pause, and then the speaker continues with a correction or repetition or simply starts over. An example of conversational data annotated with SUs (\.) and IPs (+) is given below.

yeah \. yeah \. I mean + oh it's you know +  
we're about to do like the + the uh fiesta bowl  
there \.

In this example, as in the experiments described here, we retain only the IPs associated with the end of an edit, ignoring IPs associated only with the onset of a filler. Though fillers are part of the standard DARPA scoring task, most instances are easily detected by lexical identity and so the presence of the filler IP is not as useful in language processing.

In previous work [4], we demonstrated that a state-of-the-art SU detector, relative to a pause-based segmenter, gives more than 45% of the possible error reduction in parser performance, in experiments with the structured language model [5]. Here, we show that similar results hold for two other, state-of-the-art statistical parsers. This motivates exploring a more direct integration of detected SUs and IPs into the speech recognition process via a metadata-informed language model. We describe two possible statistical frameworks, and then focus the discussion on the key issue of weakly supervised training to generate sufficient data for language model training.

**Table 1.** WER and SER for the RT-04F metadata test set the MDE system based on the SRI-only STT output.

STT system	WER	SER $\pm$ SU	SER w/ type
reference	0.0	26.70	37.30
SRI-only	18.7	40.51	51.09

## 2. MDE SYSTEMS

Results in this work span a year’s time period, so they involve two different SU/IP detection systems with results reported using two different scoring tools. The error measure in all cases is a slot error rate (SER) similar to the WER metric utilized by the speech recognition community, i.e. dividing the combined number of SU insertions, deletions and (if SU type is scored) substitutions by the total number of reference SUs. Error figures are given mainly to highlight the differences in simple vs. state-of-the-art detection results. Since the earlier results do not count type errors, the later results are reported both with and without type substitution errors.

In the parsing work, the system used here for SU and IP detection is the same as described in [2], modulo slight differences in training data. It combines decision tree models of prosody with a hidden event language model in a hidden Markov model (HMM) framework for detecting events at each word boundary, similar to [6]. Differences include the use of lexical pattern matching features (sequential matching of words or POS tags) as well as prosody cues in the decision tree, and having a joint representation of SU and IP boundary events rather than separate detectors. On the DARPA RT-03F metadata test set [7] using an earlier version of the annotation specifications [8], the model has 35.0% slot error rate (SER) for the binary SU distinction (75.7% recall, 87.7% precision), and 68.8% SER for edit IPs (41.8% recall, 79.8% precision) on reference transcripts, using the `rt_eval` scoring tool.<sup>1</sup> While these error rates are relatively high, it is a difficult task, and the SU performance was among the best in the 2003 NIST evaluation.

In the language modeling work, we use the more recent MDE system described in [3]. SU error rates for the reference transcripts and the STT output on the recent evaluation set are reported in Table 1, based on the `md-eval` (v18) scoring tool. This system is improved relative to the previous work, though the results are not directly comparable because of the difference in test sets, scoring tools, and in annotation specifications.

<sup>1</sup>Note that the IP performance figures are not comparable to those in the DARPA evaluation, since we restrict the focus to IPs associated with edit disfluencies.

## 3. MDE AND PARSING

Parsing speech can be useful for a number of tasks, including information extraction and question answering from audio transcripts. However, parsing conversational speech presents a different set of challenges than parsing text: sentence boundaries are not well-defined, punctuation is absent, and disfluencies (edits and restarts) impact the structure of language.

Early work in parsing conversational speech was rule-based and limited in domain [9, 10]. Results from another rule-based system [11] suggested that standard parsers can be used to identify speech repairs in conversational speech. Work in statistically parsing conversational speech [12] examined the performance of a parser that removes edit regions in an earlier step, as well as more integrated edit detection and parsing [13].

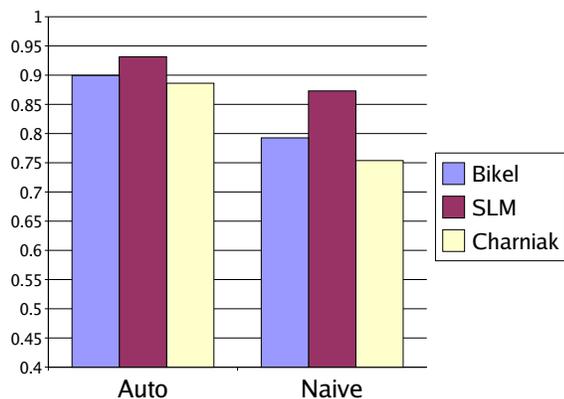
In contrast, we train a parser on the complete (human-specified) segmentation, with edit-regions included. We choose to work with all of the words within edit regions anticipating that making the parallel syntactic structures of the edit region available to the parser can improve its performance in identifying that structure. Our work also differs from prior work incorporating prosodic cues in parsing conversational speech, e.g. [14], in that our focus is primarily on sentence-level segmentation. While utterance-level segmentation may be reasonable to assume in simple human-computer dialog systems, it is not realistically available in recognized conversational speech. Secondly, we also look at the usefulness of sentence-internal IPs in parsing.

In prior work [4], we used the structured language model (SLM) as a parser with simple pause-based segmentation and automatically detected SUs, showing a significant improvement in parsing performance when using the automatic SUs. The data used in this work is the treebank (TB3) portion of the Switchboard corpus of conversational telephone speech, which includes sentence boundaries as well as the reparandum and interruption point of disfluencies. The data consists of 816 hand-transcribed conversation sides (566K words), of which we reserve 128 conversation sides (61K words) for evaluation testing according to the 1993 NIST evaluation choices.

Here, we confirm these findings with results on the same segmentation alternatives but using two state-of-the-art statistical parsers trained on conversational speech: the Bikel<sup>2</sup> [15] and Charniak<sup>3</sup> [12] parsers. Figure 1 shows the relative performance of each parser for the two different segmentations, comparing to the oracle performance for each using hand-annotated SUs. In order to have a single number for

<sup>2</sup>Version 0.9.9, <http://www.cis.upenn.edu/~dbikel/download.html>. For this work, we trained the Bikel parser on the Switchboard treebank parses with the Collins settings.

<sup>3</sup>Aug 2004, <ftp://ftp.cs.brown.edu/pub/nlparser/>



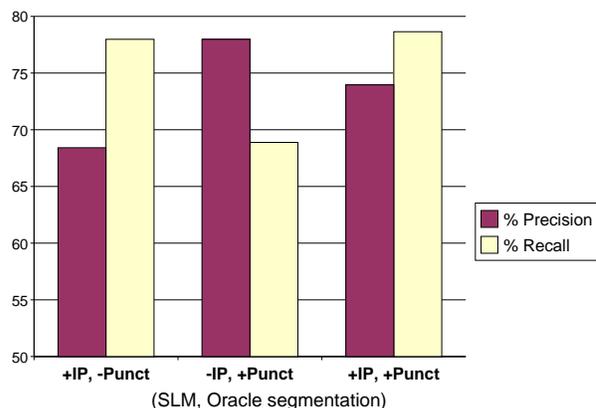
**Fig. 1.** Relative performance of different parsers for the RT03 MDE SU detection system and a simple pause-based system, in each case comparing the F-measure to that obtained with oracle SUs.

performance, we use the F-measure calculated from bracket precision and recall. (Trends with individual measures and with bracket crossing statistics follow the same patterns.) For all three parsers, there is a significant gain in performance due to using the explicit SU detection algorithm, and we would expect further gain using the current improved MDE systems. Also important is the observation that the impact of increased SU error is not lessened by using a better parser: the best oracle results were achieved with the Bikel parser, but it was much more sensitive to SU errors than the SLM.

In trying to understand the role of IPs in parsing, we ran several experiments, including some with and without human-annotated punctuation. Without punctuation, there is a small increase in parsing performance of the SLM using IPs when the SUs are automatically detected, though we have not confirmed these gains with other parsers. Including punctuation shows an interesting trade-off of bracket precision and recall using the two different cues, as illustrated in Figure 2. We see the improved precision associated with using both punctuation and IPs as evidence that sub-sentence prosodic constituents might be useful.

#### 4. METADATA-INFORMED LANGUAGE MODELS

Early work in language modeling with sentence boundaries showed the potential for improved STT performance on conversational speech [16]. Our plan is to build on this work, but update it with a richer metadata representation and the more sophisticated methods for integrating acoustic and language cues now available in current MDE systems. In addition, with advances in STT systems, it is also critical to train our language models on vastly larger corpora in or-



**Fig. 2.** Precision-recall trade-offs associated with using punctuation, oracle IPs or both in addition to SUs in parsing.

der to add value over current state-of-the-art LMs. A major problem in using the boundary events that we explored in the parsing work is that these are perceptual categories, and the cost of annotating a large corpus of speech transcripts with these categories is prohibitive.

To address this problem, we have automatically annotated the entire Fisher corpus with SU and IP boundary events. In order to reduce computational costs, the MDE system is a simplified version of [3] that does not use fundamental frequency cues, which tend to have only a small impact on performance (in CTS tasks) but a significant cost, due to the general purpose tools now used for that stage of processing. In addition, the simpler HMM-based system is used (i.e. omitting the maximum entropy model and system combination stages). Despite the use of reference transcripts, the error rate of this model is relatively high, roughly 27% SER for  $\pm$ SU detection. However, if this error was computed at the word level, i.e. counting the null events in the score, the error rate would be below 5%. In addition, the use of these detected events within a statistical framework should ameliorate the problem of errors. In fact, as we found in experiments in edit and filler detection as a second stage to SU/IP prediction [2], it is better to train the second stage on automatically detected boundary events than on hand-labeled events, since that is closer to what is available in the actual recognition process.

Given the annotated Fisher data, we are exploring two metadata-informed language models using SUs and IPs in the word stream, including a variable n-gram as a baseline (to capture boundary events as words without losing the standard 4-gram word context) and the structured language model, as in [4]. In future work, we plan to use the boundary events as conditioning factors rather than inserting them as “words”. In all cases, the language models will

not be a replacement for existing language models, because they will not be trained on all the sources used in a state-of-the-art system, which include web text and news broadcasts that are not well matched in style to the boundary events we aim to represent. Instead, the metadata-informed language model would be used in a final stage of N-best or lattice rescoring as an additional knowledge source.

While there are as yet no language modeling results to demonstrate the viability of this approach, we have obtained encouraging results combining syntactic and acoustic cues in a weakly supervised training of a model for detecting prosodic phrase and hesitation boundaries (commonly called “breaks”) in a related project. In this work, we used a small corpus (roughly 4 hours) of prosodically labeled speech from the Switchboard corpus [17], which was partitioned into test and training subsets. On the training set, we designed a decision tree classifier that combined syntactic (from Treebank) and acoustic cues to detect three classes of word boundaries: intonational phrase boundary, hesitation break and default word boundary (in the ToBI system, these corresponds to: 4, (1p,2p) and other categories combined). Given this initial model, we iteratively annotated and redesigned a decision tree using an EM approach, i.e. training the decision tree with weighted counts. We refer to this as weakly supervised because of the availability of a small amount of hand-labeled data and the Treebank parses. The model converges after only a few iterations, and the resulting labels are used to design a prosody break predictor to be used in a parsing or SLM system. Using this approach led to a relative reduction in classification error of 15% compared to using only the small hand-labeled training set. The final simplified system (no syntactic cues) obtained an accuracy of 85.8%. We also investigated a co-training approach (splitting up predictors based on acoustic and syntactic cues) and Viterbi-style iterative training, but obtained the best results with EM. Note that the EM technique extends easily to more complex detection frameworks, such as the hidden event model [18].

## 5. SUMMARY

In summary, this paper has presented evidence that the use of automatically detected boundary events improves performance of statistical parsers, which demonstrates usefulness of MDE for downstream language processing applications. In addition, the results indicate the potential for using these events in language modeling to improve speech recognition performance, though methods for weakly supervised training (i.e. avoiding hand annotation of large amounts of data) are critical to this work. Initial results in weakly supervised training of prosodic constituent phrase break models indicate that EM methods are useful for this task. In addition, the prosodic constituents themselves may be useful in lan-

guage modeling. However, the general framework is as yet untested in STT.

## Acknowledgments

This work is supported in part by DARPA contract no. MDA972-02-C-0038 and in part by NSF grant no. IIS085940. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these agencies. Thanks to J. Kim and Y. Liu for providing MDE detection systems used in this work.

## 6. REFERENCES

- [1] S. Strassel, *Simple Metadata Annotation Specification V6.2*, Linguistic Data Consortium, 2004.
- [2] J. Kim, S. E. Schwarm, and M. Ostendorf, “Detecting structural metadata with decision trees and transformation-based learning,” in *Proc. HLT-NAACL*, 2004.
- [3] Y. Liu *et al.*, “The ICSI-SRI-UW metadata extraction system,” in *Proc. ICSLP*, 2004, vol. I, pp. 577–580.
- [4] J. Kahn, M. Ostendorf, and C. Chelba, “Parsing conversational speech using acoustic segmentation,” in *Proc. HLT-NAACL, comp. vol.*, 2004, pp. 125–128.
- [5] C. Chelba and F. Jelinek, “Structured language modeling,” *Computer Speech and Language*, vol. 14, no. 4, pp. 283–332, October 2000.
- [6] Y. Liu, E. Shriberg, and A. Stolcke, “Automatic disfluency identification in conversational speech using multiple knowledge sources,” in *Proc. Eurospeech*, 2003, vol. 1, pp. 957–960.
- [7] NIST, “Rich Transcription Fall 2003 Evaluation (RT-03F),” Tech. Rep., <http://www.nist.gov/speech/tests/rt/rt2003/fall/>, 2003.
- [8] S. Strassel, *Simple Metadata Annotation Specification V5.0*, Linguistic Data Consortium, 2003.
- [9] D. Hindle, “Deterministic parsing of syntactic non-fluencies,” in *Proc. ACL*, 1983, pp. 123–128.
- [10] L. Mayfield, M. Gavalda, Y. Seo, B. Suhm, W. Ward, and A. Waibel, “Parsing real input in JANUS: a concept-based approach,” in *Proc. TMI 95*, 1995.
- [11] M. Core and K. Schubert, “Speech repairs: A parsing perspective,” in *Satellite Meeting ICPHS 99*, 1999.
- [12] E. Charniak and M. Johnson, “Edit detection and parsing for transcribed speech,” in *Proc. 2nd NAACL*, 2001, pp. 118–126.
- [13] M. Johnson and E. Charniak, “A TAG-based noisy channel model of speech repairs,” in *Proc. ACL*, 2004, pp. 33–39.
- [14] M. Gregory, M. Johnson, and E. Charniak, “Sentence-internal prosody does not help parsing the way punctuation does,” in *Proc. HLT-NAACL*, 2004, pp. 81–88.

- [15] D. Bikel, *On the Parameter Space of Lexicalized Statistical Parsing Models*, Ph.D. thesis, University of Pennsylvania, 2004.
- [16] A. Stolcke, “Modeling linguistic segment and turn boundaries for n-best rescoring of spontaneous speech,” in *Proc. Eurospeech*, 1997, pp. 2779–2782.
- [17] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, B. Byrne, and L. Carmichael, “A prosodically labeled database of spontaneous speech,” in *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 119–121.
- [18] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu, “Automatic detection of sentence boundaries and disfluencies based on recognized words,” in *Proc. ICSLP*, 1998, vol. 5, pp. 2247–2250.