

Evaluating Hierarchical Discourse Segmentation of Expository Speech

2006 Apr 08

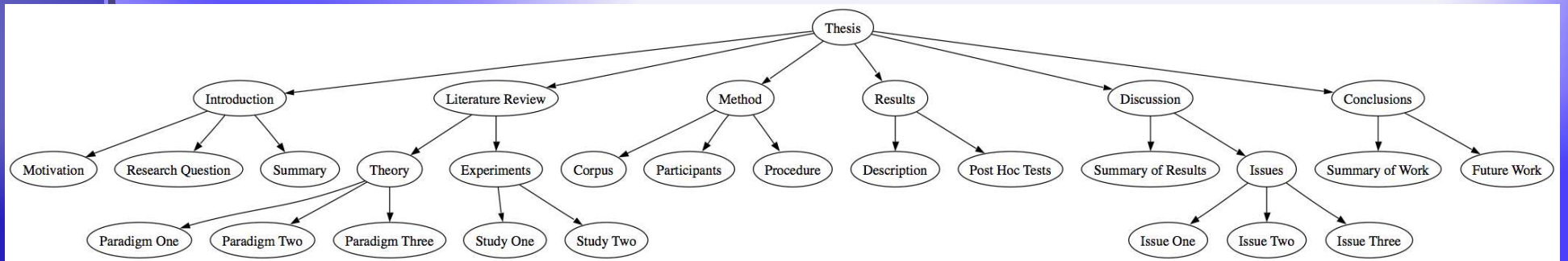
Annual LSA Colloquium at SDSU

Lucien Carroll

What is Discourse Segmentation?

- It's like finding sensible paragraph breaks and/or section breaks, but not just in expository writing
- It's shallow parsing of discourse structure
- Discourse structure:
 - Answers the question: how do people organize language beyond the sentence?
 - Is described in terms of: schemas (e.g. story grammars), informational and interactional purposes, topics and subtopics, logical or rational relations, functions of cohesive devices

Example: A Thesis



- Discourse parsing seeks the whole structure starting with individual clauses
- Linear discourse segmentation seeks the bracketing for just one level of structure
- Hierarchical discourse segmentation seeks the bracketing corresponding to just the tree above

What is it good for?

- Assists in
 - Automatic summarization (Angheluta et al., 2002)
 - Information retrieval (Kaszkiel & Zobel, 1997)
 - Anaphora resolution (Walker, 1997)
 - Question answering (Chai & Jin, 2004)
- Helps test theories of discourse structure (Passonneau & Litman, 1997)
- Enriches linguistic corpora (Csomay, 2005)

Overview

- Look at three holes in discourse segmentation research
 - Lack of diversity in genres
 - Little work in hierarchical segmentation algorithms
 - No work in quantitative measures for HDS
- Describe what I've done to address them
- Examine how well this fulfills the need
- Discuss what more is needed

Hole One: Genre Diversity

- Discourse structure theory has been interested in conversation, spoken and written narrative, and expository text
- Empirical discourse segmentation research has focused overwhelmingly on news data, some expository text and narrative, little of anything else
 - News data has been the most available data
 - Evaluating on speech data is hard because it doesn't come with headlines or headings to use as evaluation reference (a “gold standard”)

Hole Two: Hierarchical Segmentation

- Discourse structure theories are explicitly hierarchical (many theories resemble syntactic structures)
- Discourse parsing is hierarchical but has focused on small scale—a few hundred words or smaller
- Linear segmentation (i.e. like paragraph breaks alone) addresses large scale structure but isn't hierarchical
- Only three algorithms for hierarchical segmentation are published

Hole Three: Evaluating Hierarchical Segmentation

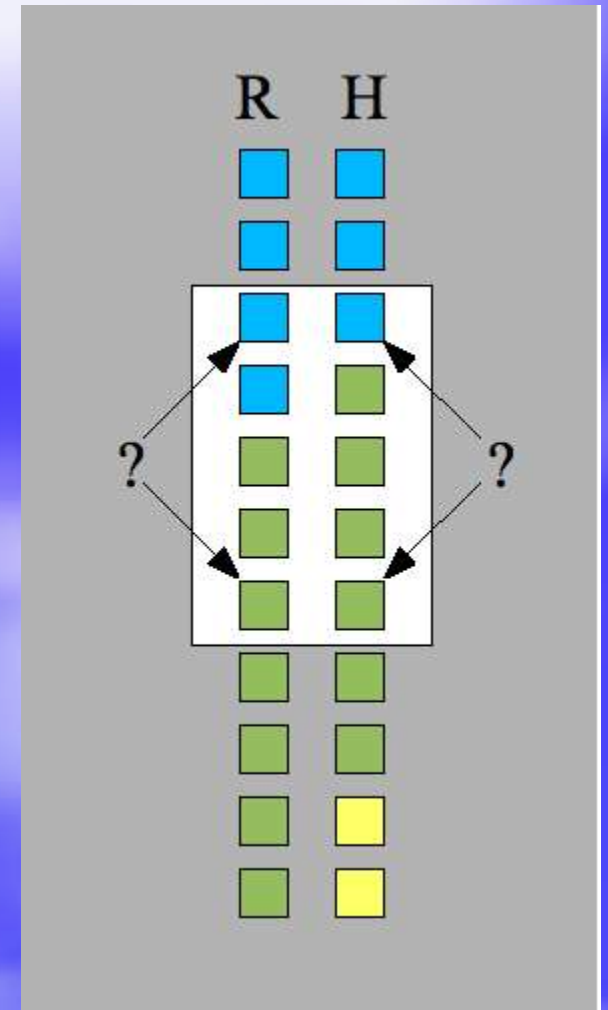
- Discourse segments are fuzzy: imprecise boundaries and debatable number
- Traditional measures like precision, recall, and crossing brackets overpenalize small errors in bracket placement
- Beeferman et al. (1999) and Pevzner & Hearst (2002) proposed probabilistic error measures to allow for variation in linear segmentation
- No comparable error measure has been offered for hierarchical segmentation

Approach

- Generalize the probabilistic error measures from linear segmentation for hierarchical segmentation
- Take two algorithms for linear segmentation that use hierarchical clustering and alter them to produce hierarchical segmentation
- Evaluate the algorithms on monologic lectures (i.e. expository speech); derive gold standard based on Passonneau & Litman (1997)

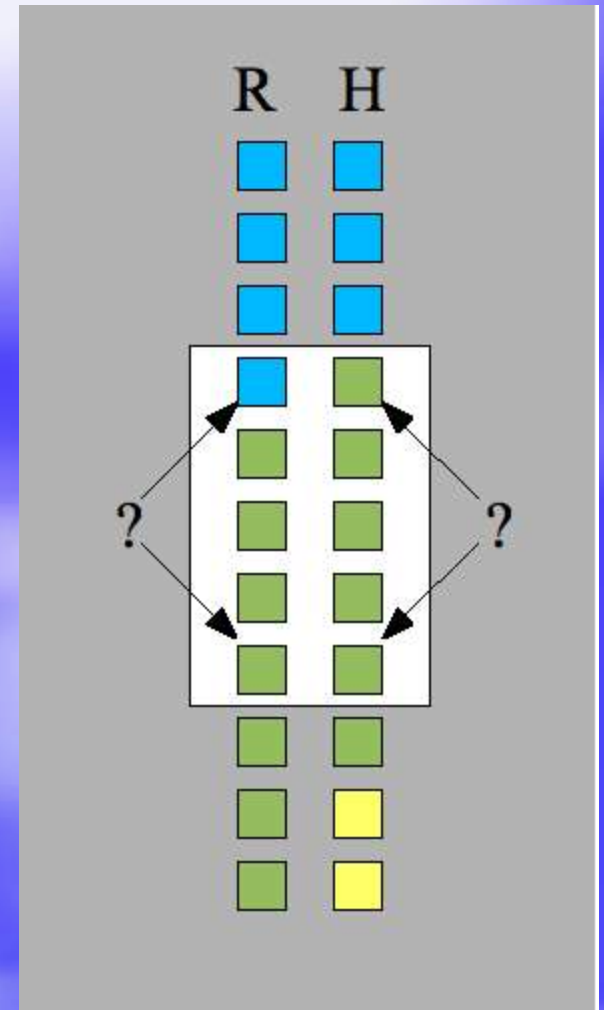
Linear Error Measure

- P_k (Beeferman et al.)
 - Looks through window at reference and hypothesized segs (width k is half average segment length in ref. seg.)
 - Asks if they agree about ends of window being in same segment
- WindowDiff (Pevzner & Hearst)
 - Asks if they agree about number of boundaries in window



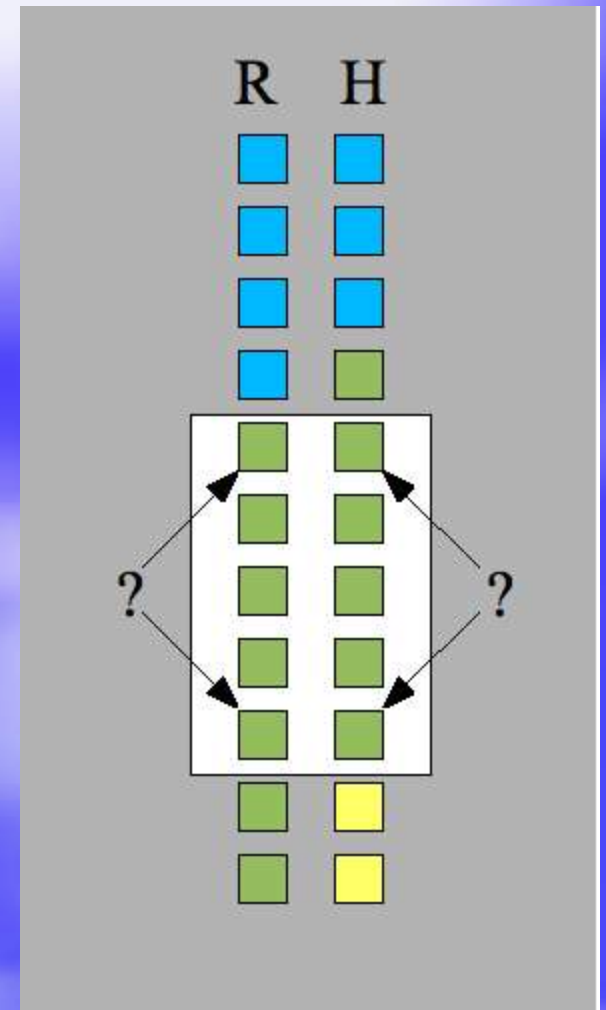
Linear Error Measure

- P_k (Beeferman et al.)
 - Looks through window at reference and hypothesized segs (width k is half average segment length in ref. seg.)
 - Asks if they agree about ends of window being in same segment
- WindowDiff (Pevzner & Hearst)
 - Asks if they agree about number of boundaries in window



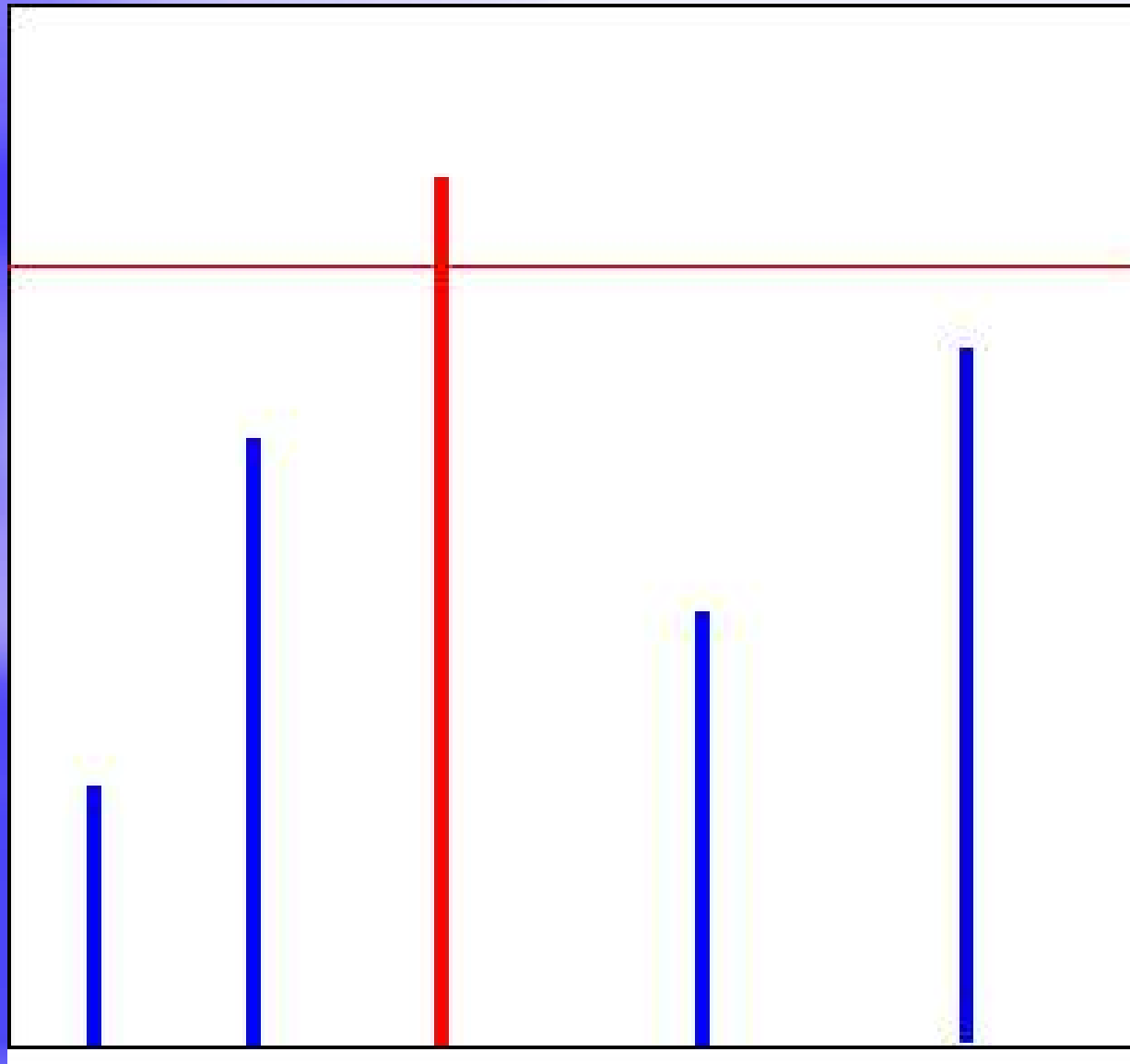
Linear Error Measure

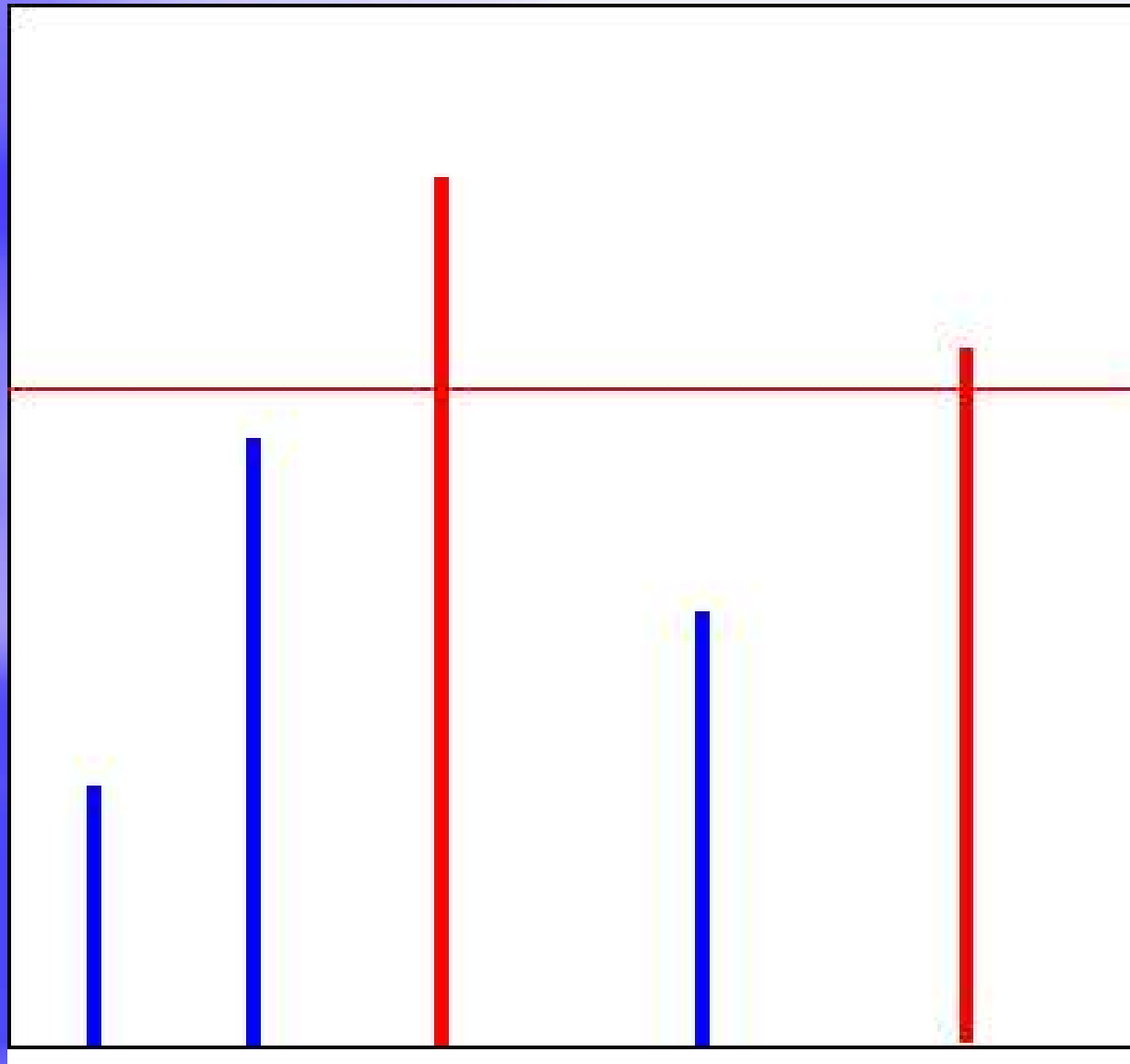
- P_k (Beeferman et al.)
 - Looks through window at reference and hypothesized segs (width k is half average segment length in ref. seg.)
 - Asks if they agree about ends of window being in same segment
- WindowDiff (Pevzner & Hearst)
 - Asks if they agree about number of boundaries in window

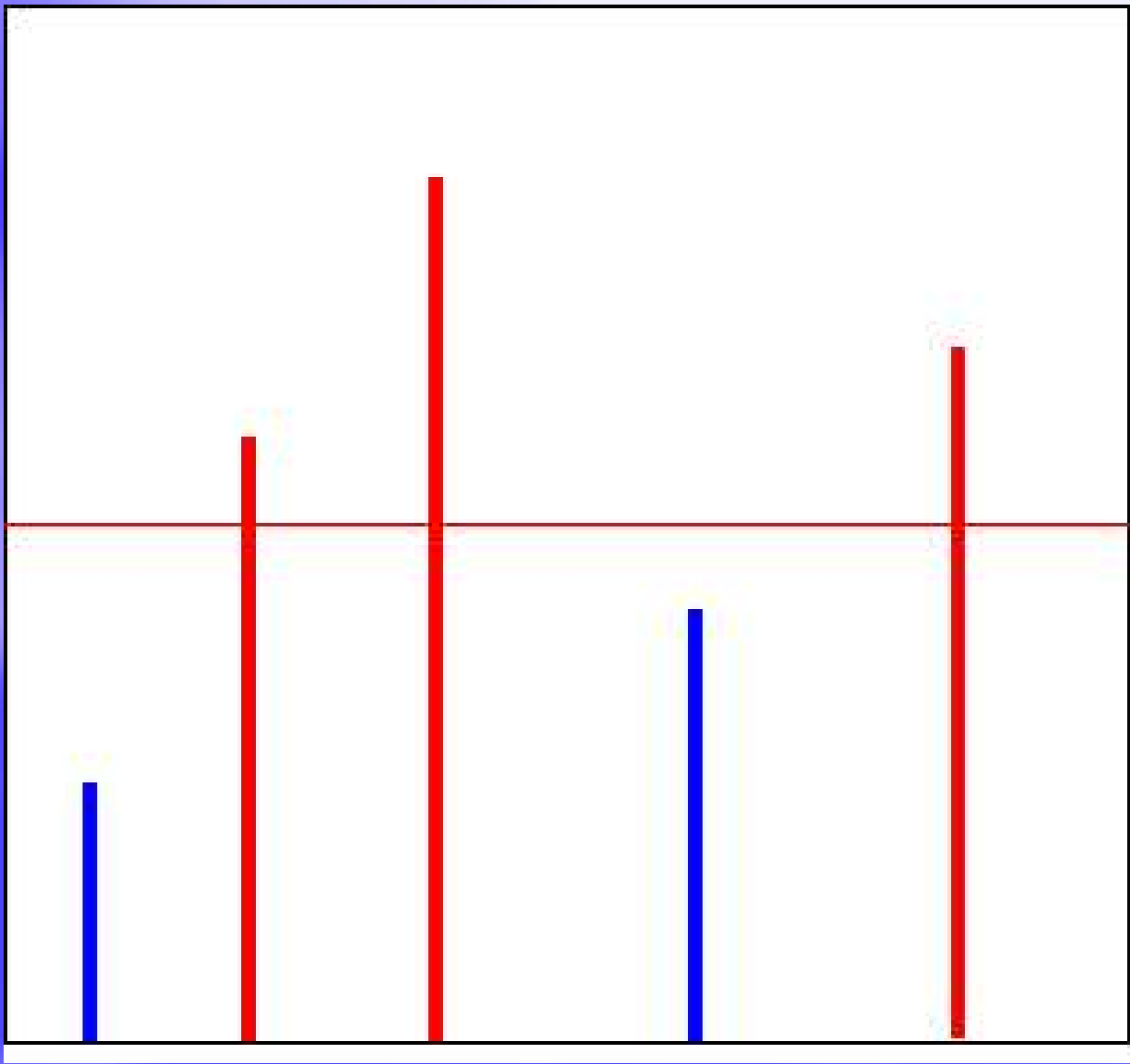


Hierarchical Error Measure

- Average P_k /WindowDiff over best linearizations of tree
 - Take n highest ranked boundaries, for n between 1 and total number of boundaries in gold standard
 - Calculate linear error for each one and average them
 - When ranks aren't all unique, average over representative sampling at each rank level







Segmentation Algorithms

- HC99: Hierarchical version of C99 (Choi, 2000)
- HCWM: Hierarchical version of CWM (Choi et al., 2001)
- NONE: Baseline with no boundaries
- BIN: Baseline made by recursive bisection
- RAND: Baseline made by a random process

Reference Segmentation

- Corpus:
 - Portions of the 7 most monologic lectures in MiCASE
 - Pauses assumed to delimit prosodic phrases
- Procedure:
 - Based on that of Passonneau & Litman (1997)
 - Untrained readers mark topic changes
 - Statistical significance of interannotator agreement identifies reference boundaries

In the following text, please check off all boxes where a "paragraph break" should go (where the topic changes).

S1: please remember that there 's another field trip coming up this Sunday if you 'd
1 like to go you could sign up with Larry Henderson by the end of the week .

2 also um ,

3 in the ,

coming events category Lorna Simpson is going to be speaking at the University
4 Museum of Art tomorrow night .

5 and she 's a contemporary um ,

6 photography conceptual artist who 's now moving into video ,

7 and is very articulate and interesting .

8 so ,

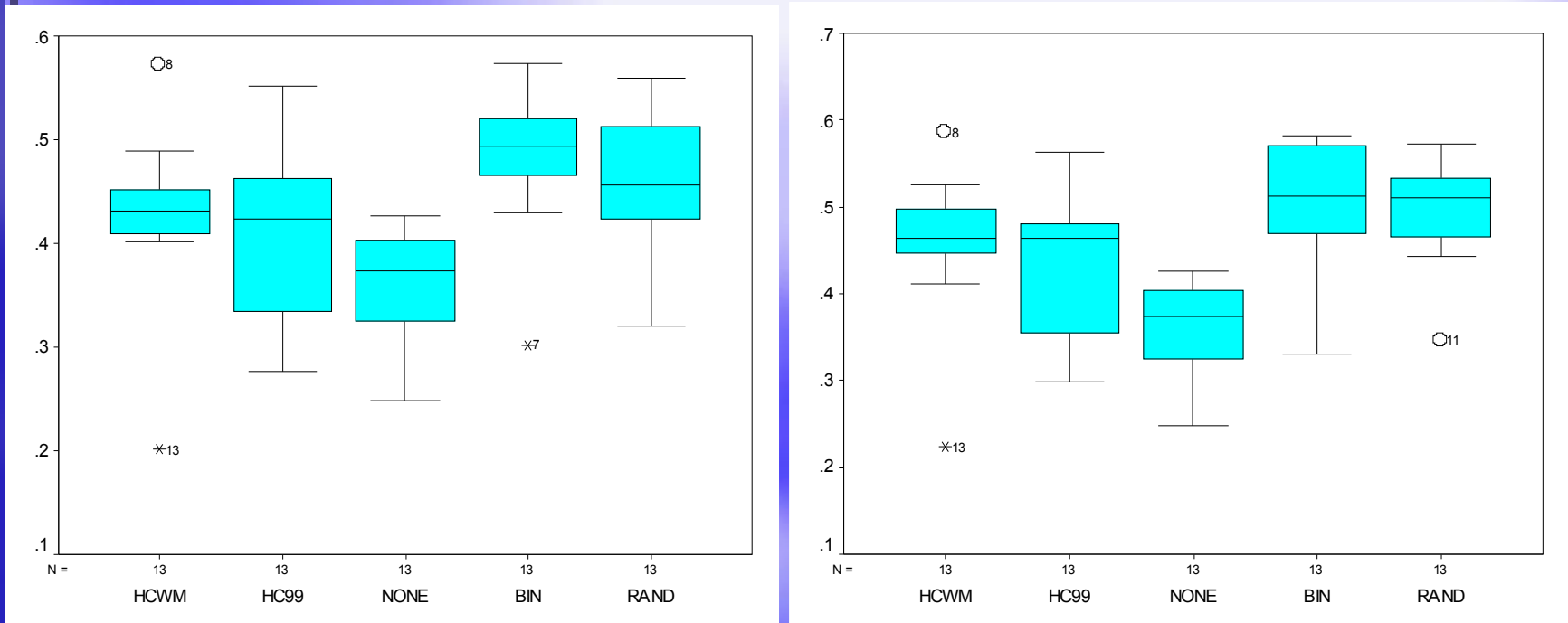
since we 've been spending all term on dead artists here 's a chance to hear a living
9 one .

10 and her presentation i think is at seven thirty _

11 is there a question ?

One page of the web form annotation interface

Evaluation 1



- Stat. sig: matched-pair two-tailed t-tests; $p < .05$
- In P_k , s.s.d.s are: NONE $<$ all, and HC99 $<$ BIN
- In WD, s.s.d.s are: NONE $<$ all, and HC99 $<$ BIN, RAND

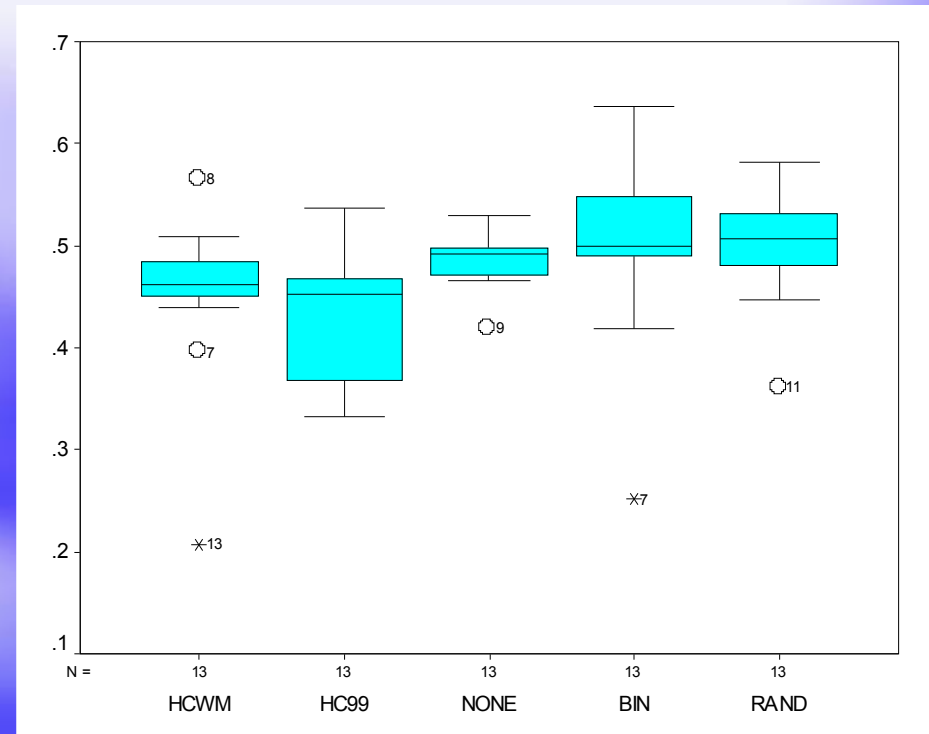
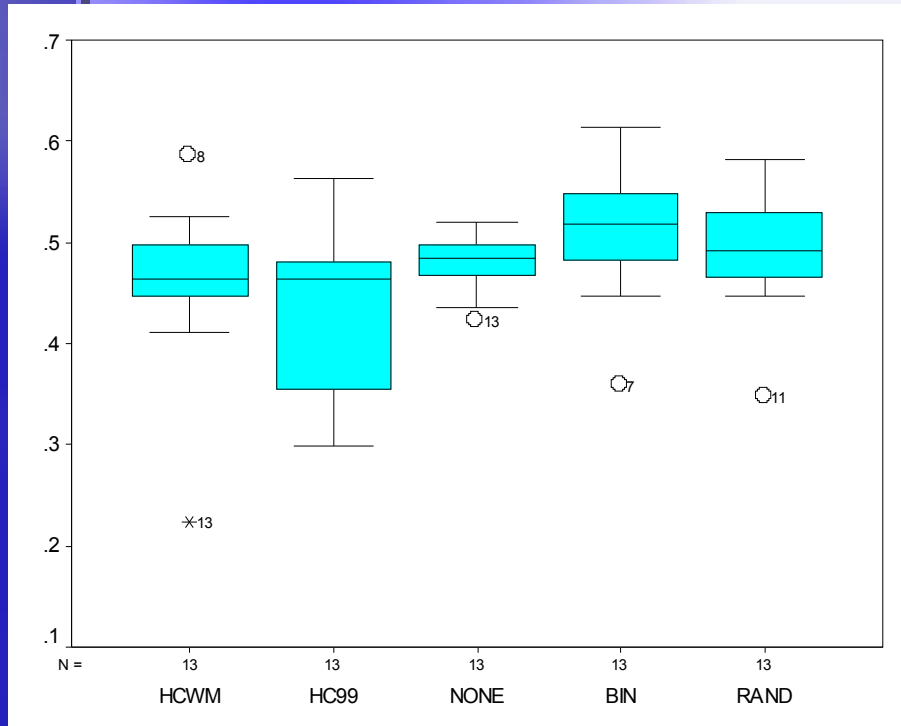
Analysis

- Why is NONE judged with such low error?
 - Few boundaries in the reference seg, not evenly spaced
 - In this situation, the false positives produced by proposing a random boundary are generally more than the false negatives removed
 - The way P_k and WD are defined, boundaries near the head and tail of discourse are undercounted

Redefine Error Measure

- Fill: In context of hierarchical segmentation, “not a discourse boundary” really means, “lower importance than what's been marked”
- Wrap: Instead of stopping at end of text, wrap window end-point around to beginning

Evaluation 2



- **WD+Fill: HC99 < NONE, BIN**
- **WD+Fill+Wrap: HC99 < NONE, RAND**

Discussion

- Appears that WD+Fill(+Wrap) is a good measure
 - High discriminating power (s.d.d.s)
 - All baselines are close to .50
 - Algorithms appear better than baselines
- Algorithms perform poorly on this genre

Conclusions

- How well does this work fulfill the need?
 - Error Measure
 - Theoretically motivated
 - Based on corresponding best measure from linear seg
 - Distinguishes algorithms from baselines (it seems)
 - Segmentation Algorithms
 - These algorithms are not much good in this genre
 - Gold Standard
 - An alternative to trained annotators
 - Method is complex: should compare to trained annotators

Further Questions

- How do HC99 and HCWM compare to trained human annotators?
- How do human annotators compare?
- How do HC99 and HCWM compare to the other algorithms in the literature?
- How can algorithms be developed for handling non-news genres better?
- This is explicitly tree structure, whereas some theories favor DAGs or general graphs.

Supplementary Slides

- Algorithm Implementation
- Reference Segmentation Procedure
- References

Algorithm Implementation

- C99 (Choi, 2000) uses hierarchical clustering on a matrix of lexical vector similarities
- CWM (Choi et al., 2001) uses latent semantic indexing (LSI) vectors instead
- C99 was released for educational use as Java code, so to create HC99 I just altered Choi's code to make it keep track of the order of cluster splits
- For HCWM, I made a Java interface to an Infomap database. I tried a MiCASE LSI model, but a BNC model performed better.

Passonneau & Litman

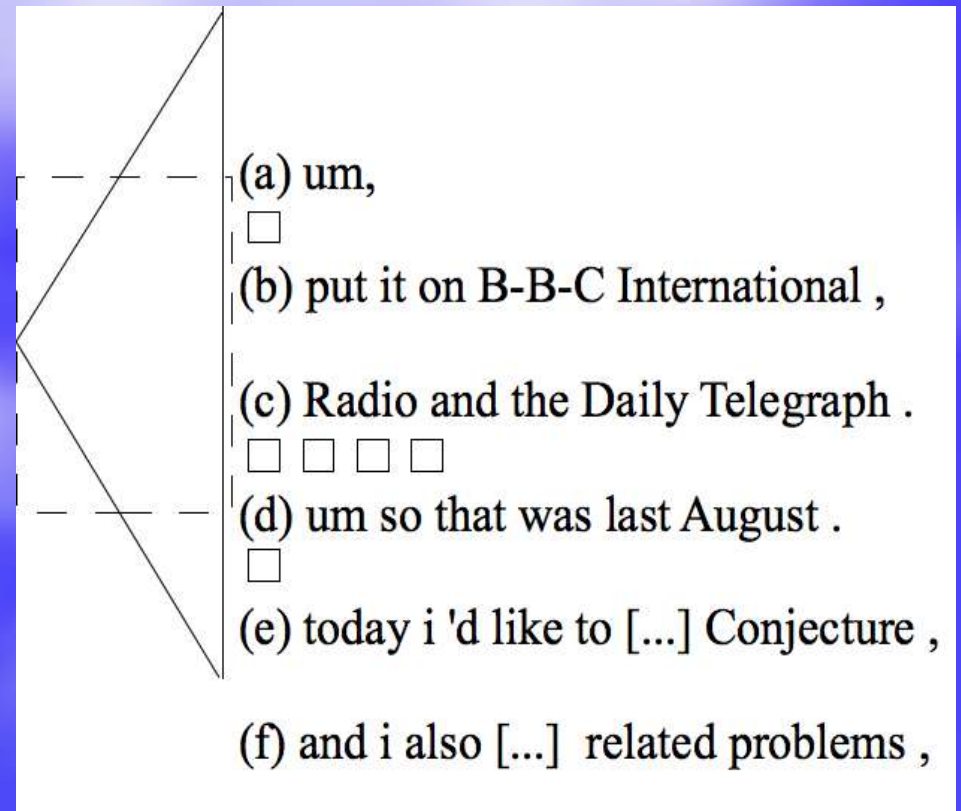
- Creating linear segmentation gold standard for Pear Narratives
- 7 untrained readers mark segment boundaries at any prosodic boundaries they like
- More than 3 readers in agreement is stat. sig.
- P&L note that
 - Readers often disagree by just one or two prosodic phrases
 - Readers differ widely about number of segments

Annotation

- Corpus:
 - Selected the seven most monologic lectures from MiCASE (highest word/turn ratios)
 - Each split into two or three sections of c.4000 words
 - One “prosodic phrase” per line (marked by punctuation in MiCASE)
- Participants: Native speakers of English, recruited from intro Ling classes, c.6 per sample
- Procedure: In internet web form, identify points where “topic changes” occur

Resolving Differences in Boundary Location

- Use a moving window for more flexible agreement criterion
- What shape?
 - Square is simple but has problems
 - Triangle behaves better
 - Gaussian is better theoretical model



Resolving Differences in Segment Count

- When some readers mark just a few boundaries and others mark many, what does that mean?
- Probably, fewer boundaries represents major branchings; many boundaries represents minor branchings
- Weight boundary marks by giving each annotator equal weight, divided evenly among that annotator's marks

Statistical Significance of Reference Boundaries

- Complex agreement criterion makes theoretical calculation of statistical significance difficult
- Calculate random agreement distribution empirically for each sample
 - Use same number of annotators and same numbers of boundary marks, but random placement
 - All agreement levels greater than minimum mark's weight go into distribution
 - All agreement levels with $p < .05$ go in gold standard, with higher stat. sig. as higher branchings

References

- Angheluta, R., Busser, R. D., and Moens, M.-F. (2002). The use of topic segmentation for automatic summarization. In *DUC 2002*.
- Beeferman, D., Berger, A., and Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.
- Csomay, E. (2005). XXXLexical EpisodesXXX. In *Discourse MDA book*.
- Chai, J. Y. and Jin, R. (2004). Discourse structure for context question answering. In *HLT-NAACL 2004 Workshop on Pragmatics of Question Answering*, pages 23–30.
- Choi, F., Wiemer-Hastings, P., and Moore, J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of 6th EMNLP*, pages 109–117.
- Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of NAACL-00*, pages 26–33.
- Kaszkiel, M. and Zobel, J. (1997). Passage retrieval revisited. In *Proceedings of 20th ACM SIGIR*, pages 178–185.
- Passonneau, R. J. and Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.
- Pevzner, L. and Hearst, M. (2001). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 16(1).
- Simpson, R. C., Briggs, S. L., Ovens, J., and Swales, J. M. (2000). The Michigan corpus of academic spoken English.
- Walker, M. A. (1997). *Centering in Discourse*, chapter Centering, Anaphora Resolution, and Discourse Structure. Oxford University Press.