Evaluating Hierarchical Discourse Segmentation

Lucien Carroll
lucien@ling.ucsd.edu

UC San Diego
Mentors: Eniko Csomay (SDSU), Rob Malouf (SDSU), Andy Kehler (UCSD)

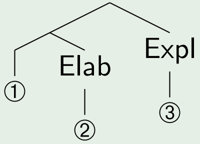
NAACL10
June 4, 2010

Why discourse segmentation?

The clauses in a coherent discourse have grouping relationships.

Example

1. John bicycled downtown.
2. It took an hour.
3. He really likes donuts.

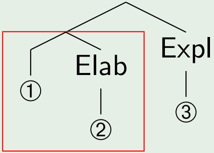


Why discourse segmentation?

The clauses in a coherent discourse have grouping relationships.

Example

1. John bicycled downtown.
2. It took an hour.
3. He really likes donuts.

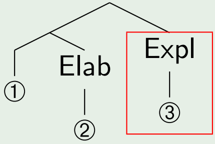


Why discourse segmentation?

The clauses in a coherent discourse have grouping relationships.

Example

1. John bicycled downtown.
2. It took an hour.
3. He really likes donuts.

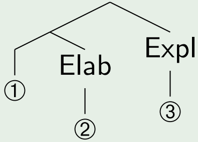


Why discourse segmentation?

The clauses in a coherent discourse have grouping relationships.

Example

1. John bicycled downtown.
2. It took an hour.
3. He really likes donuts.



Discourse segmentation

Identification of the boundaries between groups of sentences

Relevant to summarization, information retrieval, coreference resolution, genre analysis, etc.

Summary

A proposed segmentation error measure is an improvement for linear segmentation and works well for hierarchical segmentation.

What is the structure of discourse?

Is discourse made of general graphs, DAGs, trees or sequences?

- Discourse structure theory: trees, DAGS, or general graphs
- Discourse parsing: trees or DAGS
- Segmentation: sequences (mostly)

The sequence model is useful, but it's missing important structure

What is discourse segmentation?

The canonical outline has a tree structure

Thesis Outline

1. Introduction
   - Lit Review
     - Field 1
     - Field 2
   - Question
2. Methods
   - Corpus
   - Procedure
3. Results
   - Summary
   - Discussion
     - Issue 1
     - Issue 2
4. Conclusions
   - Summary
   - Further work

What is discourse segmentation?

1	Thesis									
2	Intro			Methods		Results			Conclusions	
3	Lit		Q	Corpus	Proc	Sum	Discussion		Sum	Future
4	F1	F2					Q1	Q2		

What is discourse segmentation?

1	Thesis									
2	Intro			Methods		Results			Conclusions	
3	Lit		Q	Corpus	Proc	Sum	Discussion		Sum	Future
4	F1	F2					Q1	Q2		

Linear discourse segmentation

Find the boundaries for one level of structure

Hierarchical discourse segmentation

Find the boundaries for each level of structure

What is discourse segmentation?

1	Thesis									
2	Intro			Methods		Results			Conclusions	
3	Lit		Q	Corpus	Proc	Sum	Discussion		Sum	Future
4	F1	F2					Q1	Q2		

Linear discourse segmentation

Find the boundaries for one level of structure

Hierarchical discourse segmentation

Find the boundaries for each level of structure

Hierarchical segmentation algorithms

A few previous studies have explored hierarchical segmentation algorithms, but evaluation is problematic.

Study	Evaluation Method
(Yaari, 1997)	as one linear segmentation
(Slaney & Ponceleon, 2001)	visual comparison against outline
(Angheluta et al., 2002)	evaluation of summarization system
(Eisenstein, 2009)	two levels of linear segmentation

The evaluation problem

Evaluation of linear segmentation is already problematic.

- Even trained experts disagree about number and placement of boundaries
- A small difference in boundary placement is essentially still agreement

Example

...
It took three hours
He really likes donuts
And that place is good
Once, I got there early
...

The evaluation problem

Evaluation of linear segmentation is already problematic.

- Even trained experts disagree about number and placement of boundaries
- A small difference in boundary placement is essentially still agreement

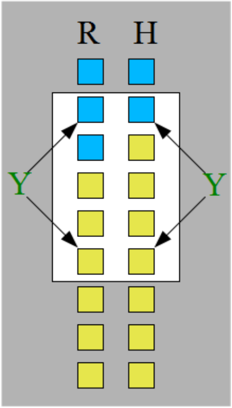
Example

...
It took three hours
He really likes donuts
And that place is good
Once, I got there early
...

The Beeferman error measure

Partially overcomes those issues.
Compares Reference (R) and Hypothesized (H) segmentations:

1. Set window width to half the average segment length (in R)
2. Run window through the text, checking if R and H agree:
   - Are there boundaries in the window?
3. Calculate error as percentage of disagreement

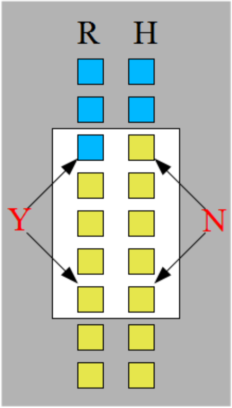


(Beeferman et al., 1999)

The Beeferman error measure

Partially overcomes those issues.
Compares Reference (R) and Hypothesized (H) segmentations:

1. Set window width to half the average segment length (in R)
2. Run window through the text, checking if R and H agree:
   - Are there boundaries in the window?
3. Calculate error as percentage of disagreement

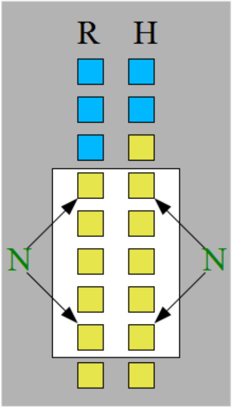


(Beeferman et al., 1999)

The Beeferman error measure

Partially overcomes those issues.
Compares Reference (R) and Hypothesized (H) segmentations:

1. Set window width to half the average segment length (in R)
2. Run window through the text, checking if R and H agree:
   - Are there boundaries in the window?
3. Calculate error as percentage of disagreement



(Beeferman et al., 1999)

Issues with the Beeferman measure

1. ***Beef*** **is more sensitive to false negatives than false positives**
   - Because two close boundaries are almost like a single boundary.
   - WindowDiff penalizes them equally.
     Instead of checking for a boundary in the window, *WD* counts number of boundaries in the window

Pevzner and Hearst (2001)

Issues with the Beeferman measure

2. **Sparse bracketing is favored**
   Recall that: $Beef(Perfect) = 0\%$, $Beef(Chance)$ *about* 50%
   - For typical segmentations, and baselines NONE and ALL: $Beef(NONE) = 45\%$, $Beef(ALL) = 55\%$
   - *WD* favors sparse bracketing even more: $WD(NONE) = 45\%$, $WD(ALL) = 99\%$

Issues with the Beeferman measure

3. **Both *Beef* and *WD* under-count boundaries near beginning and end of the text.**
   The window conventially runs from $[1, k]$ to $[N - k, N]$
4. **Neither *Beef* nor *WD* apply to hierarchical segmentations**

A hierarchical measure

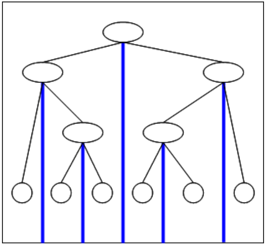
Purpose

Extend the *Beef* / *WD* error measures to evaluate hierarchical segmentation

A hierarchical measure

$Hier_{Beef}$ error is weighted average of *Beef* measurements over a series of linear segmentations

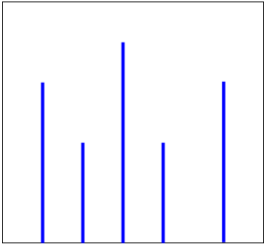
1. In the first step, only the highest boundaries are used
2. Each following step includes one more level of boundaries



A hierarchical measure

Hier$_{Beef}$ error is weighted average of *Beef* measurements over a series of linear segmentations

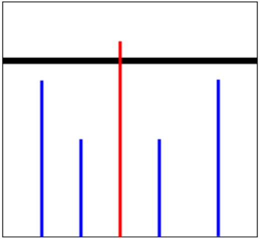
1. In the first step, only the highest boundaries are used
2. Each following step includes one more level of boundaries



A hierarchical measure

Hier$_{Beef}$ error is weighted average of *Beef* measurements over a series of linear segmentations

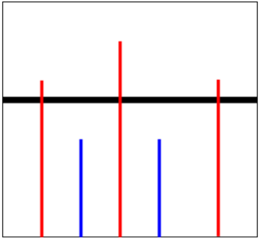
1. In the first step, only the highest boundaries are used
2. Each following step includes one more level of boundaries



A hierarchical measure

$Hier_{Beef}$ error is weighted average of *Beef* measurements over a series of linear segmentations

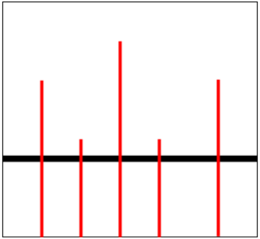
1. In the first step, only the highest boundaries are used
2. Each following step includes one more level of boundaries



A hierarchical measure

$Hier_{Beef}$ error is weighted average of *Beef* measurements over a series of linear segmentations

1. In the first step, only the highest boundaries are used
2. Each following step includes one more level of boundaries



A hierarchical measure

Hierarchical atom-error rate

Given reference (R) and hypothesized (H) boundaries:

- for prominence level i, let R_i: all boundaries of levels 1 through i

A hierarchical measure

Hierarchical atom-error rate

Given reference (R) and hypothesized (H) boundaries:

- for prominence level i, let R_i: all boundaries of levels 1 through i
- choose H_i s.t. $|H_i| = |R_i|$ and no boundary not in H_i is more prominent than any boundary in H_i (if not unique, then average over them)

A hierarchical measure

Hierarchical atom-error rate

Given reference (R) and hypothesized (H) boundaries:

- for prominence level i, let R_i: all boundaries of levels 1 through i
- choose H_i s.t. $|H_i| = |R_i|$ and no boundary not in H_i is more prominent than any boundary in H_i (if not unique, then average over them)
- weight c_i: the number of boundaries at level i in R

A hierarchical measure

Hierarchical atom-error rate

Given reference (R) and hypothesized (H) boundaries:

- for prominence level i, let R_i: all boundaries of levels 1 through i
- choose H_i s.t. $|H_i| = |R_i|$ and no boundary not in H_i is more prominent than any boundary in H_i (if not unique, then average over them)
- weight c_i: the number of boundaries at level i in R

A hierarchical measure

Hierarchical atom-error rate

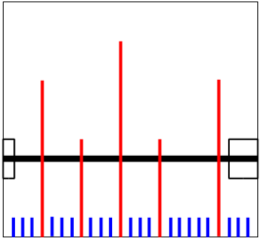
Given reference (R) and hypothesized (H) boundaries:

- for prominence level i, let R_i: all boundaries of levels 1 through i
- choose H_i s.t. $|H_i| = |R_i|$ and no boundary not in H_i is more prominent than any boundary in H_i (if not unique, then average over them)
- weight c_i: the number of boundaries at level i in R

$$Hier_{Beef} = \frac{1}{|R|} \sum_i c_i Beef(R_i, H_i)$$

A hierarchical measure

- Let the window wrap around the end of the text (from i to $(i + k) \bmod N$.)
- Possible boundaries that aren't labeled as segment boundaries in H are boundaries of least prominence
  $\hookrightarrow$ NONE and ALL equivalent



Evaluation on Choi data

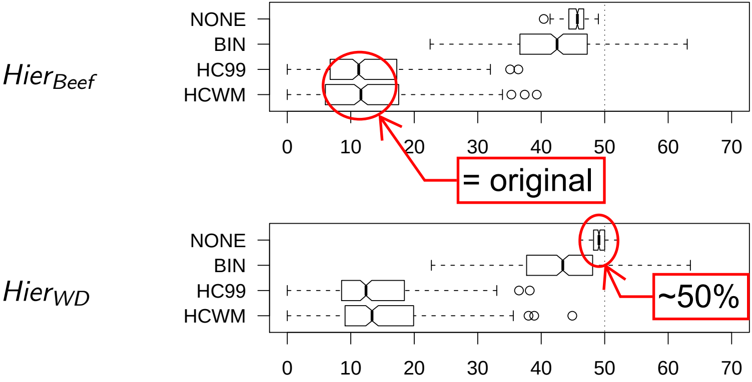
Evaluate:

- Hyp Seg: Freddy Choi's C99 and CWM modified for hierarchical output (HC99, HCWM)
- Ref Seg: Choi's standard data (concatenated text chunks)
- Baseline: BIN (recursive bisection) and NONE

Choi, Wiemer-Hastings, and Moore (2001)

No sparse bracketing preference

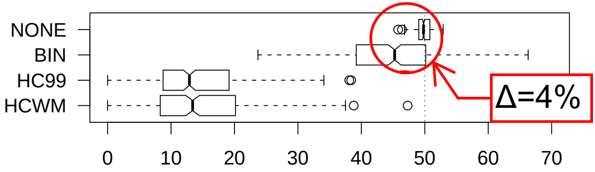
Sparse bracket preference corrected, with $Hier_{WD}(NONE) = 50\%$



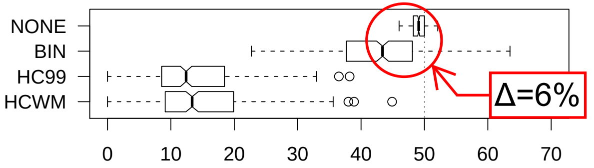
Informative about text ends

Wrapping around text end subtly informative: BIN vs. NONE

Hier$_{WD}$
without
wrapping



Hier$_{WD}$
with
wrapping



Evaluation on Wikipedia data

Evaluate:

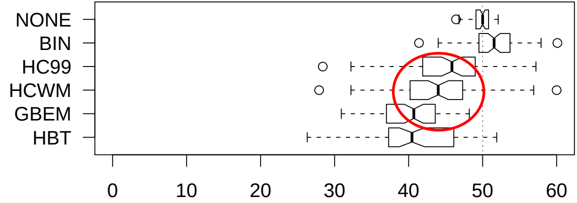
- Hyp Seg: HC99, HCWM, and Eisenstein's HierBayes-topic (HBT) and GreedyBayes-EM (GBEM)
- Ref Seg: 66 Wikipedia articles with 4 levels of headings
- Baseline: BIN and NONE

Eisenstein (2009)

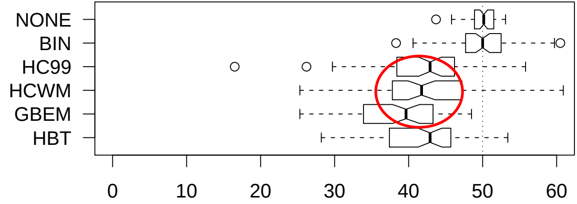
Unreliable prominences from linear seg methods

HC99, HCWM and GBEM have lower error when Ref levels ignored

Hier$_{WD}$



Hier$_{WD}$ without Ref levels



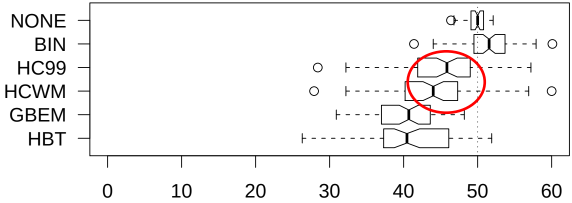
---

in 2-tail paired t-tests, $p < .0001$

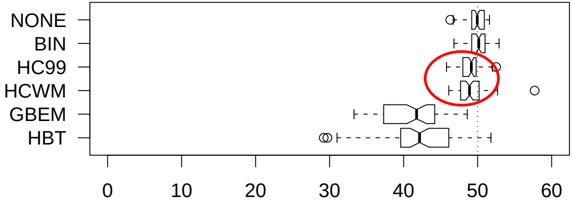
Prominent boundaries have more accurate locations

Ignoring Hyp boundary levels raises error (esp. HCWM and HC99)

Hier$_{WD}$



Hier$_{WD}$ without Hyp levels



---

in 2-tail paired t-tests, $p < .0001$

Summary

- Proposed an extension of the *Beef* / *WD* error measure for hierarchical discourse segmentation
- It is actually an improvement for linear discourse segmentation also
- It distinguishes more informed hierarchical segmentations from less informed ones
- This can guide systems that integrate discourse parsing research and linear segmentation research

Acknowledgements

Special thanks to:

- Mentors: Eniko Csomay, Rob Malouf, Andy Kehler
- Discussants: CompLing Lab, CPL Lab, reviewers, ...
- Code/data: Freddy Choi and Jacob Eisenstein

References

Angheluta, R., De Busser, R., & Moens, M.-F. (2002). The use of topic segmentation for automatic summarization. In *DUC 2002. Proceedings of the Workshop on Multi-document Summarization Evaluation at the 40th ACL.*

Beeferman, D., Berger, A., & Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, *34*(1-3), 177-210.

Choi, F., Wiemer-Hastings, P., & Moore, J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of 6th EMNLP* (p. 109-117).

Eisenstein, J. (2009). Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of NAACL09.*

Pevzner, L., & Hearst, M. (2001). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, *16*(1).

Slaney, M., & Ponceleon, D. (2001). Hierarchical segmentation: Finding changes in a text signal. *Proceedings of SIAM 2001 Text Mining Workshop*, 6-13.

Yaari, Y. (1997). Segmentation of expository texts by hierarchical agglomerative clustering. In *Proceedings of RANLP'97.*

The Beeferman error measure

Beeferman error

In a document of N atoms (words or sentences), with $N/2k$ reference segments,

$$P_k = \frac{1}{N-k} \sum_{i=1}^{N-k} \delta(\delta(r_i, r_{i+k}), \delta(h_i, h_{i+k}))$$

where arguments r_i and h_i are the segment index of atom i in the reference and hypothesized segmentations, respectively, and δ is the Kronecker delta function.

Beeferman et al. (1999)

The WindowDiff error measure

WindowDiff error

Instead of testing for a boundary in the window, count boundaries in the window.

$$WD = \frac{1}{N-k} \sum_{i=1}^{N-k} \delta(r_i - r_{i+k}, h_i - h_{i+k})$$

Pevzner and Hearst (2001)

Distribution of sentences per segment

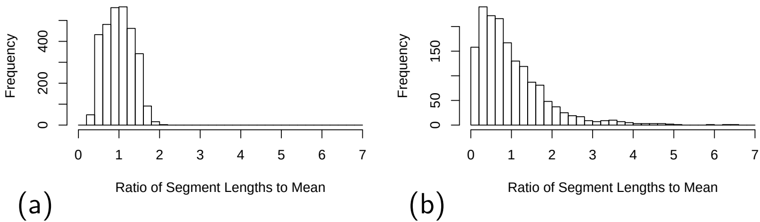


Figure: Distribution of sentences per segment for (a) Choi standard data (b) Wikipedia data