

DUC 2006 Pyramid Evaluation

Rebecca J. Passonneau


Center for Computational Learning Systems

Columbia University

Acknowledgments

- n Hoa Dang
- n Columbia University (Kathy McKeown)
- n Guideline contributors, testers (Lucy Vanderwende, Adam Goodkind, Guy LaPalme, . . .)
- n Pyramid Creators (Adam Goodkind, Sergey Sigelman, Lucy Vanderwende, Inderjeet Mani, Qui Long)
- n Participants (21 sites)

Pyramid Overview

- n Human summarizers select overlapping content
- n A pyramid represents and quantifies the overlap of Summary Content Units (SCUs) found in multiple model summaries
- n Two pyramid scores based on SCU annotations
 - u Original \approx Precision
 - u Modified \approx Recall
- n Manual annotation  reliability assessment
 - u Pyramid annotations (LREC 2006)
 - u Peer annotations (DUC 2005)

Sample SCU from D0631

[Label: The Concorde crossed the Atlantic in less than 4 hours]

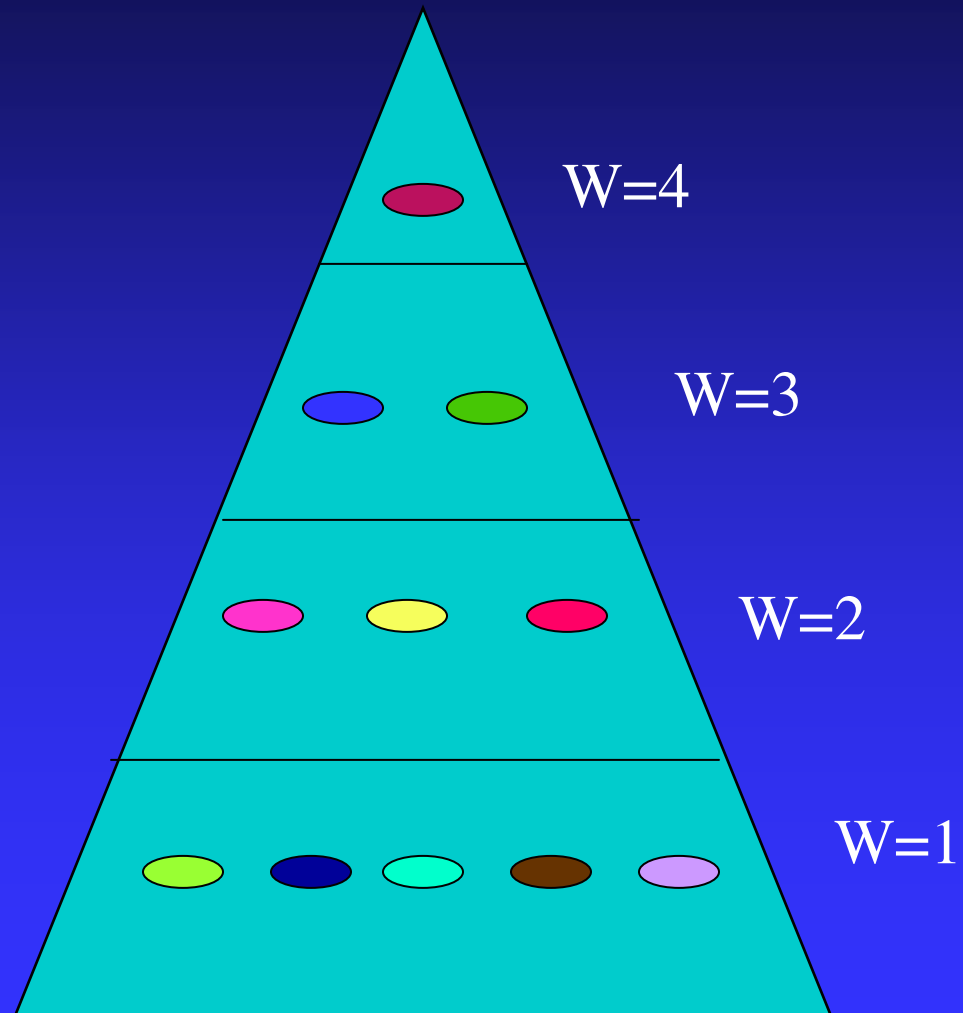
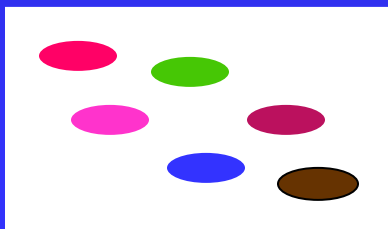
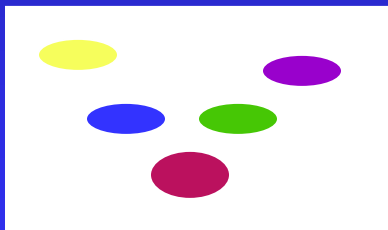
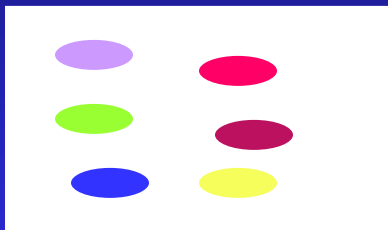
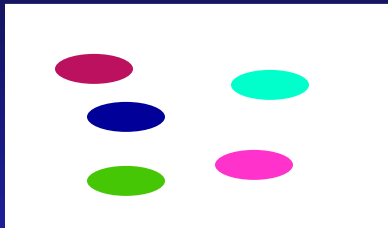
Sum1 < making the transatlantic flight in 3 and ½ hrs >

Sum2 < The Concorde could make the flight in between New York and London or Paris in less than four hours >

Sum3 < completing its journey from London to New York in about 3 hours, 30 minutes >

Sum4 < took less than 4 hrs to cross the Atlantic >

Building a Pyramid from Model Summaries (N=4)



2006 Pyramid effort

- n New version of DUCView, annotation guidelines
- n Pyramids for 20 of the document sets
 - u High clarity ratings
 - u Even distribution of assessors (summary writers)
- n Pyramid annotation
 - u 6 individuals at 3 sites, 2 with prior experience
- n Peer annotation: 21 peers plus the baseline
 - u New procedure: “peer” review
- n Only modified pyramid score (normalized to average # SCUs per model for each pyramid)

Brief Comparison with 2005

- n Same characteristics for document clusters
- n 4 instead of 7 model summaries
 - u 2005: mean of mean SCU weight = 1.9
 - u 2006: mean of mean SCU weight = 1.56
- n Possibly simpler task (cf. Litowski, DUC 2006)
- n Possibly more coherent pyramids
- n Improved systems
 - u 19/25 (76%) beat the baseline in 2005
 - u 17/21 (81%) beat the baseline in 2006

ANOVA Results

n Dependent variable: modified score

n 9 Factors:

- u **Peerid (p~0)**

- u **Setid (p~0)**

- u 5 LingQuality ratings

- u **Content responsiveness (p=0.0001)**

- u Overall responsiveness (includes readability)

System Differences (Tukey's HSD)

Peers	> peers
1, 17, 18, 25, 25 ($N=5$)	NIL
22, 29, 32 ($N=3$)	1
19, 24, 33 ($N=3$)	1, 35, 17, 18 ($N=4$)
2, 3, 6, 14, 15 ($N=5$)	1, 35, 17, 18, 25 ($N=5$)
28	1, 35, 17, 18, 25, 29 ($N=6$)
27	1, 35, 17, 18, 25, 29, 32, 22 ($N=8$)
8	1, 35, 17, 18, 25, 29, 32, 22, 14 ($N=9$)
10, 23	1, 35, 17, 18, 25, 29, 32, 22, 14, 19, 5, 33, 24, 3, 6, 2, 15 ($N=17$)

For Illustration: Group Means

Peers	Mean modified score
1, 17, 18, 25, 35 ($N=5$)	.113 ($\Delta \sim .06$)
22, 29, 32 ($N=3$)	.169
19, 24, 33 ($N=3$)	.176
2, 3, 6, 14, 15 ($N=5$)	.199
28	.205
27	.210
8	.214
10, 23	.241 ($\Delta \sim .03$)

DOCSET

Differences

Docsets	Mean pyramid score
5	.065 ($\Delta \sim .06$)
1, 3, 8, 15, 47	.133
50	.135
45, 30	.158
28	.164
16, 17, 20, 29	.172
27	.197
14	.229 ($\Delta \sim .03$)
43	.252
40	.269
24	.286
31	.357 ($\Delta \sim .07$)

Content Evaluation

- n Perfect correlation with mean pyramid score per content level

Content Assessment	Mean Pyr Score
1	.12
2	.17
3	.19
4	.21
5	.22

Comparison with DUC 2005

- n Many more significant differences among peers using Tukey
 - u 2005: 2 distinct comparison sets
 - u 2006: 8 distinct comparison sets
- n Better correlation with responsiveness
 - u 2 assessors in 2005, $r=.81$; $.90$
 - u 1 assessor in 2006, $r=1$

Factors Affecting System Scores

- n Differences in document set difficulty/coherence
- n Pyramid characteristics
 - u Mean SCU weight
 - u Pyramid size and proportion of weight 1 SCUs
- n Score variability
 - u 2005: sd = .14
 - u 2006: sd = .09
- n Better systems
 - u 2005 mean system score range: .20 to .06
 - u 2006 mean system score range: .24 to .11

Semantics of Pyramids

- n More highly weighted SCUs
 - u more general
 - u less dependent on meaning of other SCUs

Generality of Highly Weighted SCUs

n W=4

- u D0603: *Wetlands help control floods*
- u D0605: *Exercise helps arthritis*

n W=1

- u D0603: *In underdeveloped countries the increase of rice-planting has negative impacts on wetlands*
- u D0605: *Arthroscopic knee surgery appears to reduce pain, for unknown reasons*

Semantic Independence of Highly Weighted SCUs

n $W=4$

- u D0640: *The Kursk sank in the Barents Sea*
- u D0617: *Egypt Air Flight 990 crashed*

n $W=1$

- u D0640: *The escape hatch [of *] was too badly damaged to dock in 7 attempts*
- u D0617: *Tail elevators [of*] were in an uneven position, indicating a possible malfunction*

Impressions/Questions

- n Does greater difficulty of a docset correlate with greater specificity/interrelatedness?
 - u D0647 is associated with lower mean pyramid scores
 - u 9 SCUs of $W=4$ are all very specific (about sea rescue of Cuban child, Elian Gonzales)
 - u 5 of 9 SCUs of $W=4$ refer to other SCUs

Conclusion

- n Systems have improved: DUC roadmap has been successful
- n Evaluation document sets have good coverage; but can we begin to characterize document set difficulty?
- n Would pyramid scores (intrinsic) correlate with any extrinsic measures?