Cross-Domain Study of N-grams Co-Occurrence Metrics - A Case in Summarization

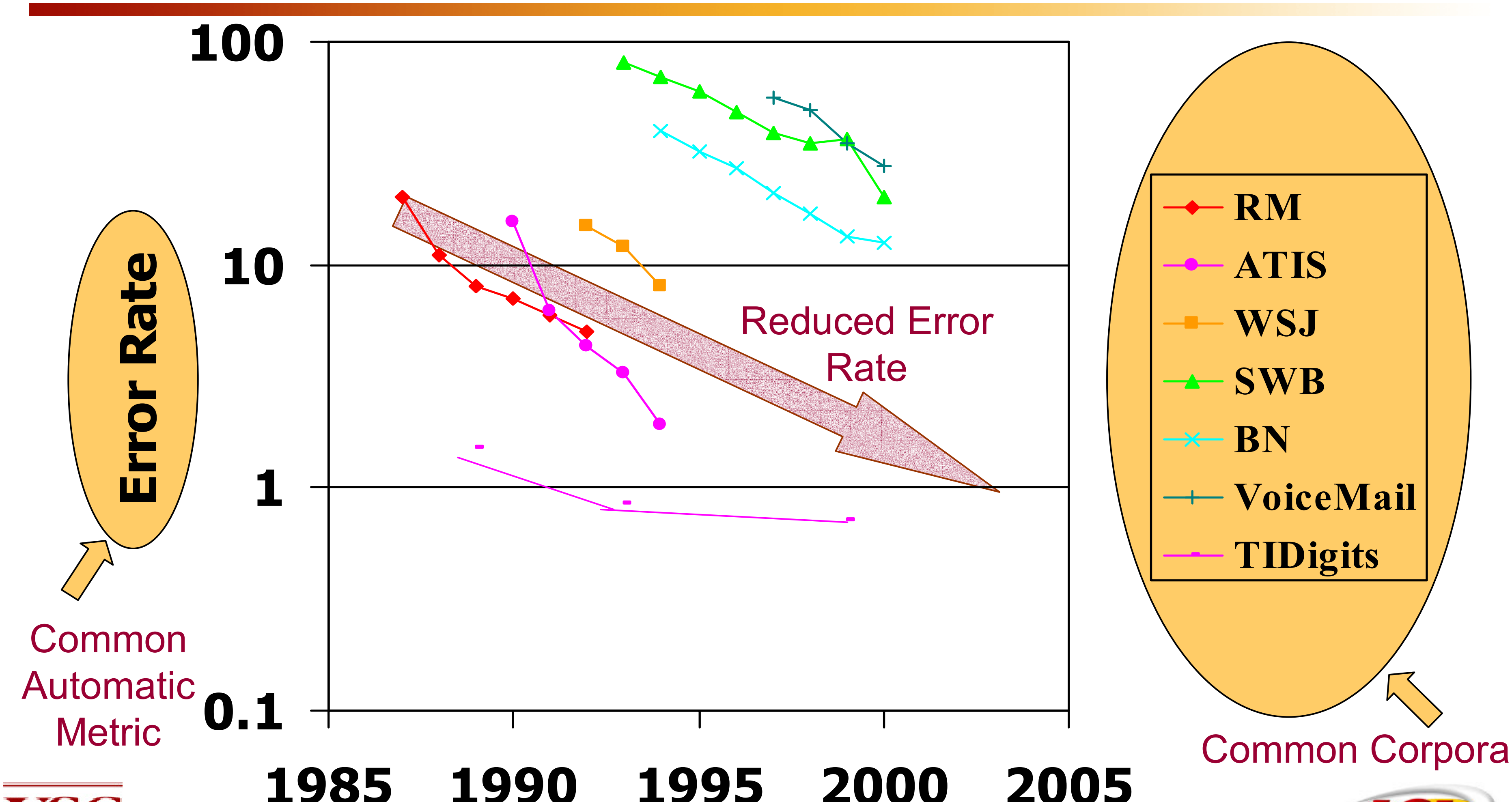
Chin-Yew Lin

Information Sciences Institute
University of Southern California
Marina del Rey, CA, USA
cyl@isi.edu

Agenda

- Motivation
- Document Understanding Conference (DUC)
- MT vs. Summarization
- How to Evaluate Evaluation Metrics
- Evaluations
- Conclusion

Progress in Speech Recognition



Mukund Padmanabhan and Michael Picheny, HLT Group, IBM Research, ASR Workshop, Paris, Sep 2000.

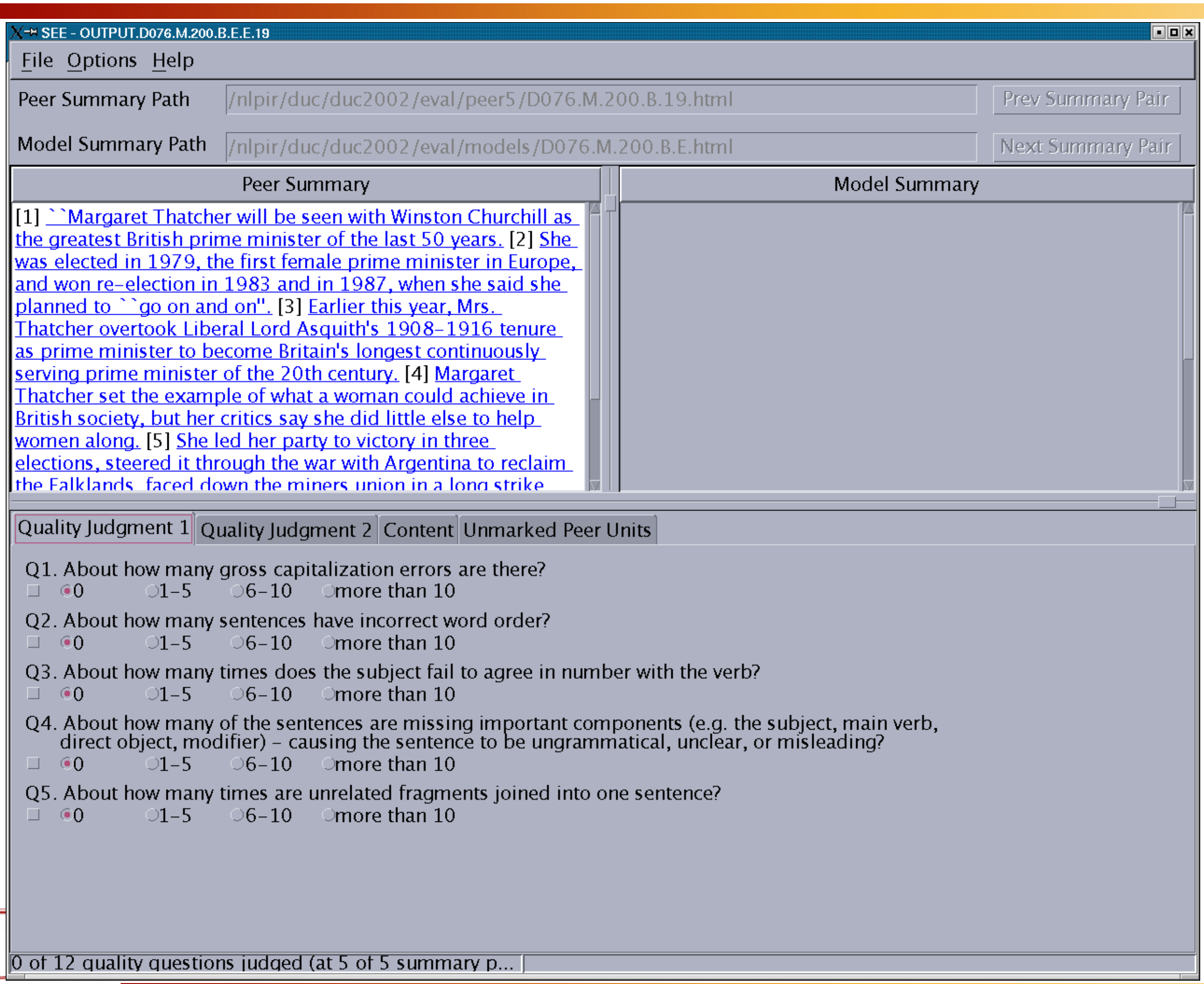
Recent Activities in Automatic Text Summarization

- DUC 2001, 2002, and 2003
  - Tasks
    - Single-doc summarization (30 topics DUC 2001 and 2002)
    - Single-doc headline generation (30 topics, DUC 2003)
    - Multi-doc summarization
      - Generic 10, 50, 100, 200 (2002) , and 400 (2001) words summaries
      - Short summaries of about 100 words in three different tasks in 2003
        - » focused by an event (30 TDT clusters)
        - » focused by a viewpoint (30 TREC clusters)
        - » in response to a question (30 TREC Novelty track clusters)
  - Participants
    - 15 systems in DUC 2001
    - 17 systems in DUC 2002
    - 21 systems in DUC 2003
  - Manual evaluations [Over 2002]
    - 3926 pairwise comparisons in DUC 2001
    - 6785 pairwise comparisons in DUC 2002

Cost of Evaluations

- 10 summarizer/evaluators for each DUC
- Time allocation
  - document selection (1-2 days)
  - manual summary (abstract/extract) creation
  - evaluation of summaries (4 hours/day)
- Total time spent in each DUC
  - 3000 hours (DUC2001: 25% in evaluation; DUC2002: 35% in evaluation)
- Can we reuse these resources and shorten the turn around time of evaluation?

SEE: Overall Candidate Quality



SEE ñ

Summary

Evaluation

Environment

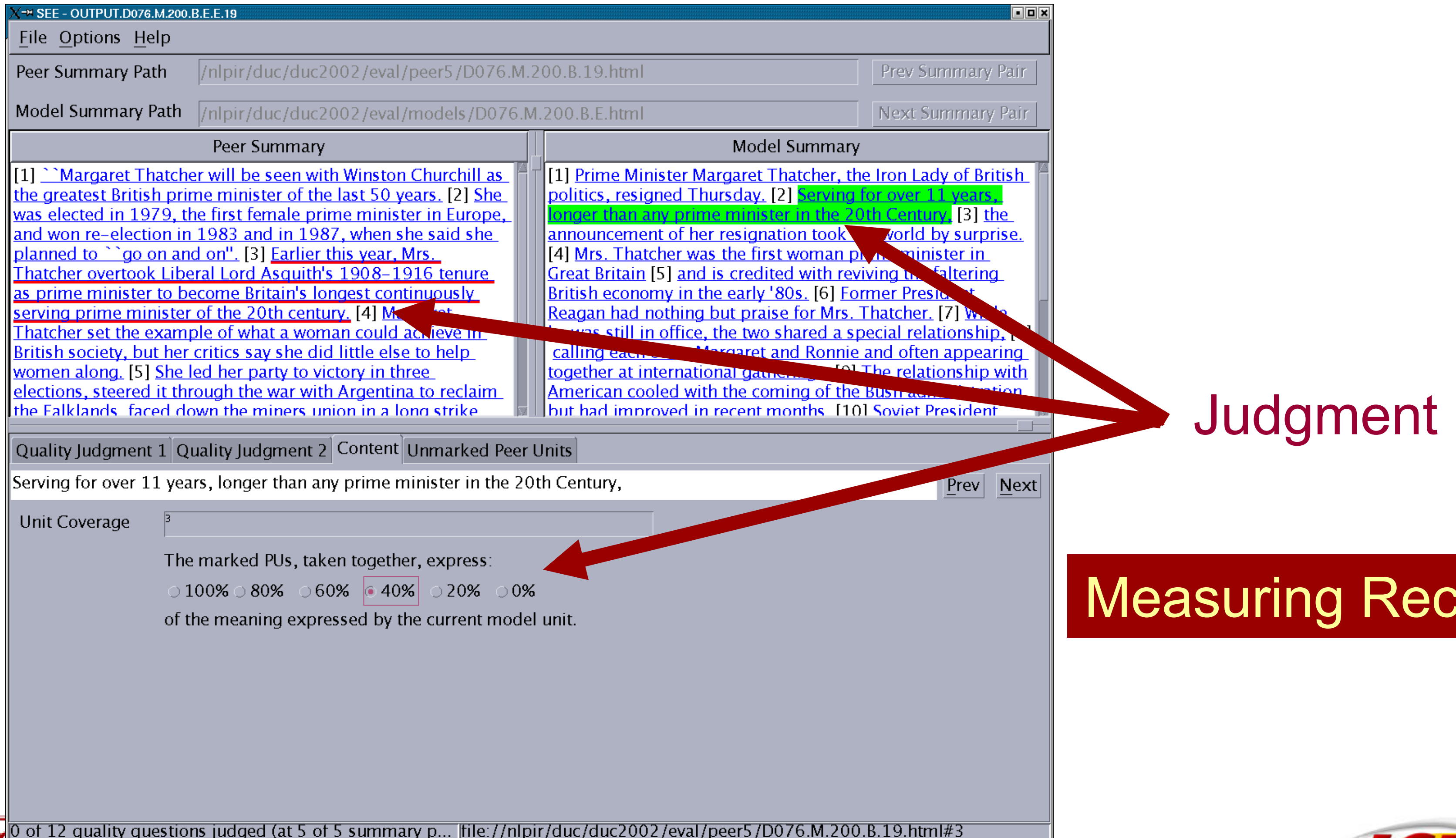
12 Questions about Candidate Quality (Q1 - Q5)

1. About how many gross capitalization errors are there?
2. About how many sentences have incorrect word order?
3. About how many times does the subject fail to agree in number with the verb?
4. About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) – causing the sentence to be ungrammatical, unclear, or misleading?
5. About many times are unrelated fragments joined into one sentence?

12 Questions about Candidate Quality (Q6 – Q12)

6. About how many times are articles (a, an, the) missing or used incorrectly?
7. About how many pronouns are there whose antecedents are incorrect, unclear, missing, or come only later?
8. For about how many nouns is it impossible to determine clearly who or what they refer to?
9. About how times should a noun or noun phrase have been replaced with a pronoun?
10. About how many dangling conjunctions are there ("and", "however"...)?
11. About many instances of unnecessarily repeated information are there?
12. About how many sentences strike you as being in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship, or just don't fit in topically with neighboring sentences?

Measuring Content Coverage



Measuring Recall!

Sample Headlines

1. Researchers using newest drugs to detect early signs of schizophrenia.
2. Early schizophrenia identification allows treatment but stigma of sufferer raises questions
3. Australian, Yale studies: medication for preschizophrenics: identify, medicate, ethical?
4. Yale, Melbourne, Zyprexa, anti-psychotic drugs, Risperdal, Cornblatt, ethicists, pre-schizophrenic, Applebaum
5. No set of indicators that can predict future illness with reliability.
6. Australian study; anti-psychotic drug Zyprexa; severe mental illness; researchers; earlier schizophrenia
7. researchers schizophrenia psychotic risk illness drugs study mcglashan subjects treatment

MT and Summarization Evaluations

- Machine Translation
  - Inputs
    - Reference translation
    - Candidate translation
  - Methods
    - Manually compare two translations in:
      - Adequacy
      - Fluency
      - Informativeness
    - Auto evaluation using:
      - BLEU/NIST scores

- Auto Summarization
  - Inputs
    - Reference summary
    - Candidate summary
  - Methods
    - Manually compare two summaries in:
      - Content overlap
      - Linguistic qualities
    - Auto evaluation?
      - ?

Anatomy of BLEU Matching Score

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Precision-based Metric!

Weighted geometric average favors longer N-gram matches

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

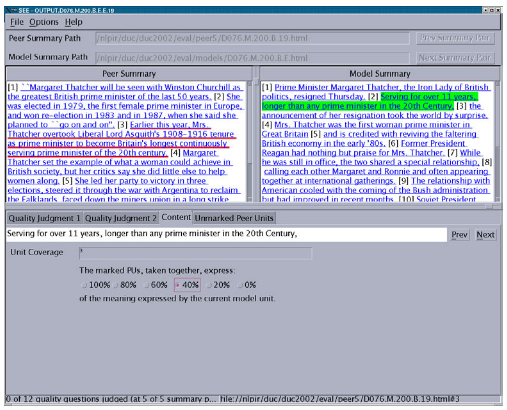
Counts of N-gram overlaps between a candidate and reference translations

$$p_n = \frac{\sum_{\mathcal{C} \in \{Candidates\}} \sum_{n\text{-}gram \in \mathcal{C}} Count_{clip}(n\text{-}gram)}{\sum_{\mathcal{C} \in \{Candidates\}} \sum_{n\text{-}gram \in \mathcal{C}} Count(n\text{-}gram)}$$

Total number of n-gram in the candidate translation

Papineni et al. 2001, ìBleu: a Method for Automatic Evaluation of Machine Translation ì IBM Research Report RC22176(W0109-022)

ROUGE: Recall-Oriented Understudy for Gisting Evaluation



ROUGEs — N-gram co-occurrence metrics measuring content overlaps

Counts of N-gram overlaps between candidate and model summaries

$$ROUGE_n = \frac{\sum_{C \in \{Model\ Units\}} \sum_{n\text{-}gram \in C} Count_{match}(n\text{-}gram)}{\sum_{C \in \{Model\ Units\}} \sum_{n\text{-}gram \in C} Count(n\text{-}gram)}$$

Total number of n-grams in the model summary

Recall-based Metric!
(fixed-length summaries)

ROUGE vs. BLEU

- ROUGE – Recall
  - Separately evaluate 1, 2, 3, and 4-grams
  - No length penalty (applying length-cutoff)
  - Verified for extraction summaries
  - Focus on content overlap
  - No correlation data for quality so far
- BLEU – Precision
  - Mixed n-grams
  - Use brevity penalty to penalize system translations that are shorter than the average reference length
  - Favors longer n-grams for grammaticality or word order

Evaluating Evaluation Metrics

- Automatic evaluation should correlate highly, positively, and consistently with human assessments. (Spearman ρ)
- The statistical significance of automatic evaluations should be a good predictor of the statistical significance of human assessments with high reliability. (R, P, F1)

➢ We verified these using DUC 2001 data. (Lin & Hovy HLT-NAACL 2003)

DUC 2001 - ROUGE(i,j) vs. Human Ranking Correlations for 4 Statistics

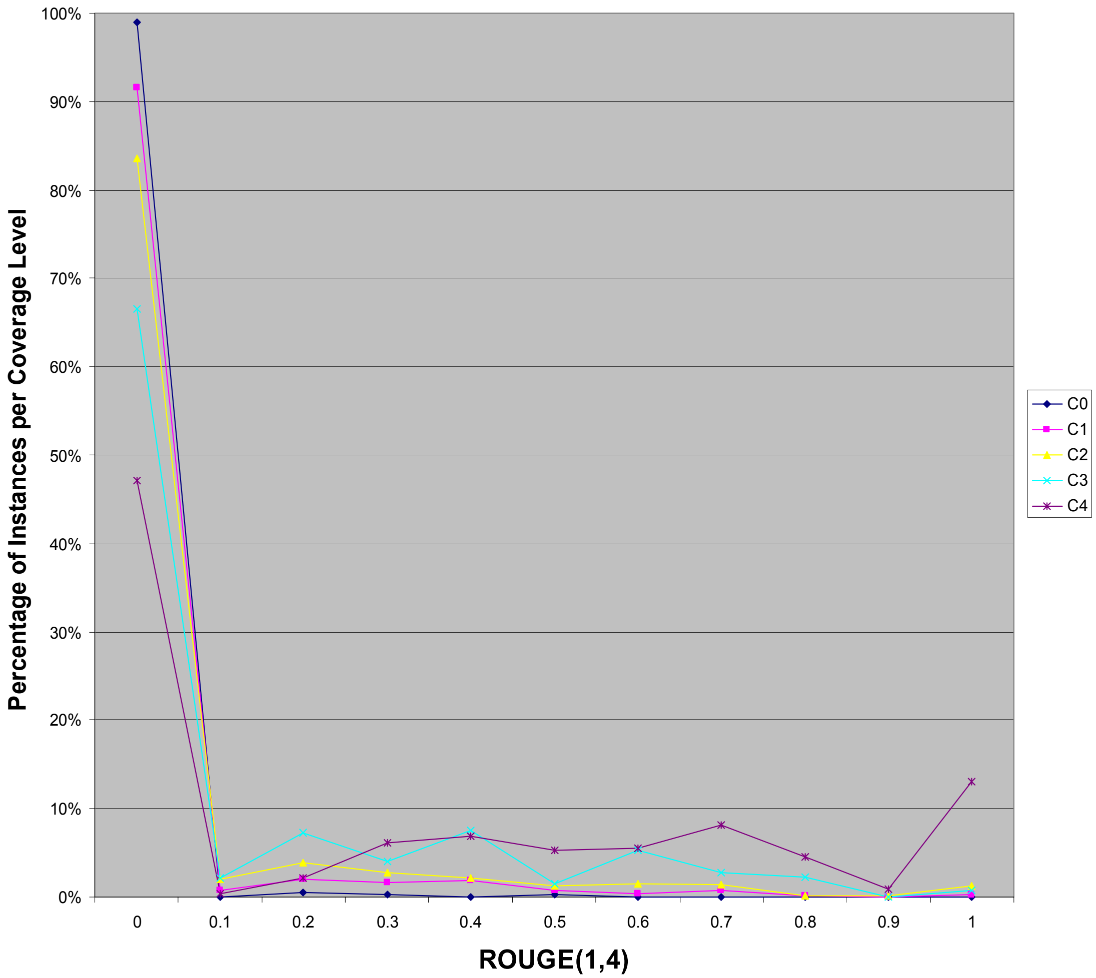
		ROUGE(1,4)	ROUGE(1,1)	ROUGE(2,2)	ROUGE(3,3)	ROUGE(4,4)
Single Doc 100	Spearman ρ	0.604	0.989	0.868	0.527	0.505
	LR_t	1.025	7.130	2.444	0.704	0.053
	Pearson ρ	0.295	0.907	0.593	0.208	0.016
	CD	0.087	0.822	0.352	0.043	0.000
Multi-Doc All	Spearman ρ	0.875	0.993	0.950	0.782	0.736
	LR_t	3.910	13.230	5.830	3.356	2.480
	Pearson ρ	0.735	0.965	0.851	0.681	0.567
	CD	0.540	0.931	0.723	0.464	0.321
Multi-Doc 50	Spearman ρ	0.546	0.879	0.746	0.496	0.343
	LR_t	2.142	5.681	3.350	2.846	2.664
	Pearson ρ	0.511	0.844	0.681	0.620	0.594
	CD	0.261	0.713	0.463	0.384	0.353
Multi-Doc 100	Spearman ρ	0.575	0.896	0.761	0.543	0.468
	LR_t	2.369	7.873	3.641	1.828	1.385
	Pearson ρ	0.549	0.909	0.711	0.452	0.359
	CD	0.301	0.827	0.505	0.204	0.129
Multi-Doc 200	Spearman ρ	0.775	0.979	0.904	0.782	0.754
	LR_t	3.243	15.648	4.929	2.772	2.126
	Pearson ρ	0.669	0.974	0.807	0.609	0.508
	CD	0.447	0.950	0.651	0.371	0.258
Multi-Doc 400	Spearman ρ	0.861	0.982	0.961	0.854	0.661
	LR_t	4.390	10.569	6.409	3.907	2.755
	Pearson ρ	0.773	0.946	0.872	0.735	0.607
	CD	0.597	0.896	0.760	0.540	0.369

LR: Linear regression *t*-test
CD: Coefficient of determination

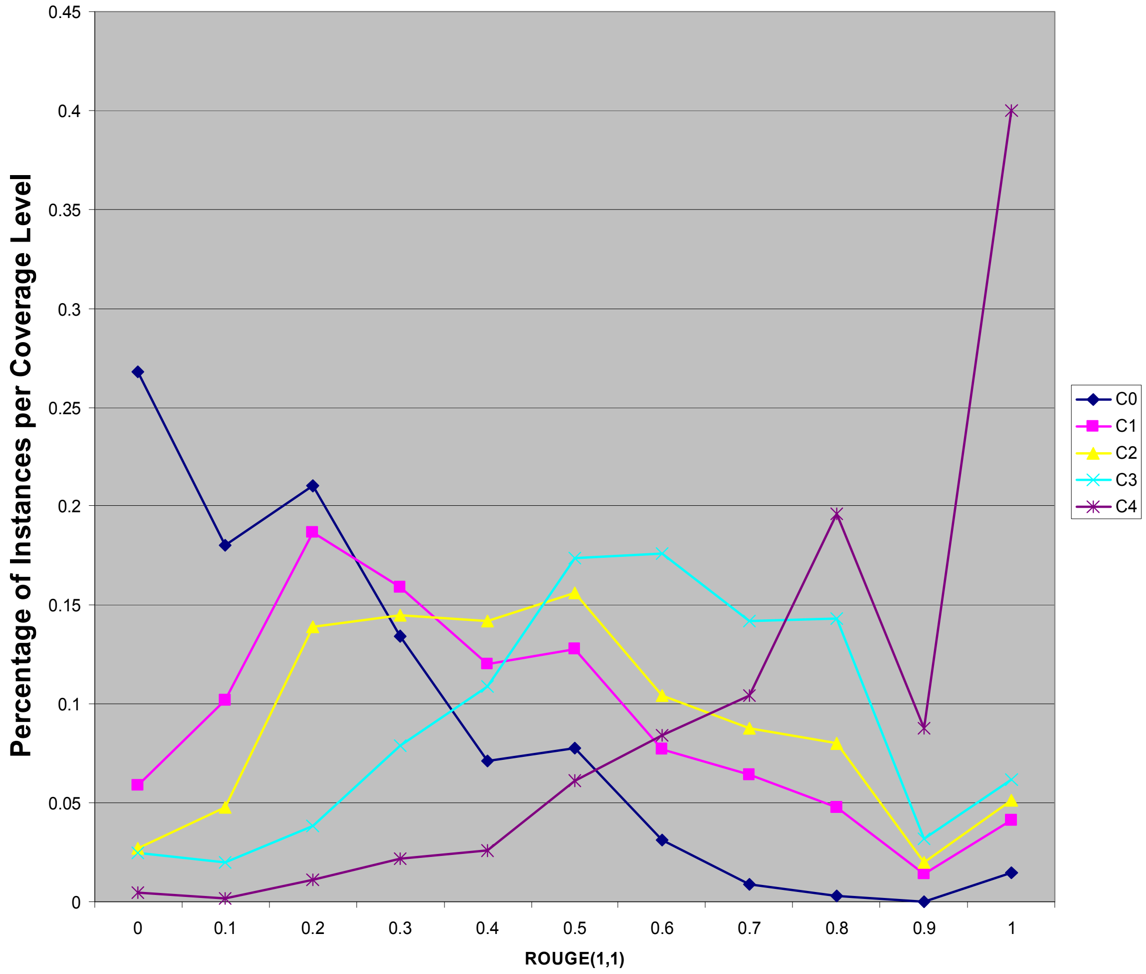
USC SCHOOL OF ENGINEERING

Information Sciences Institute

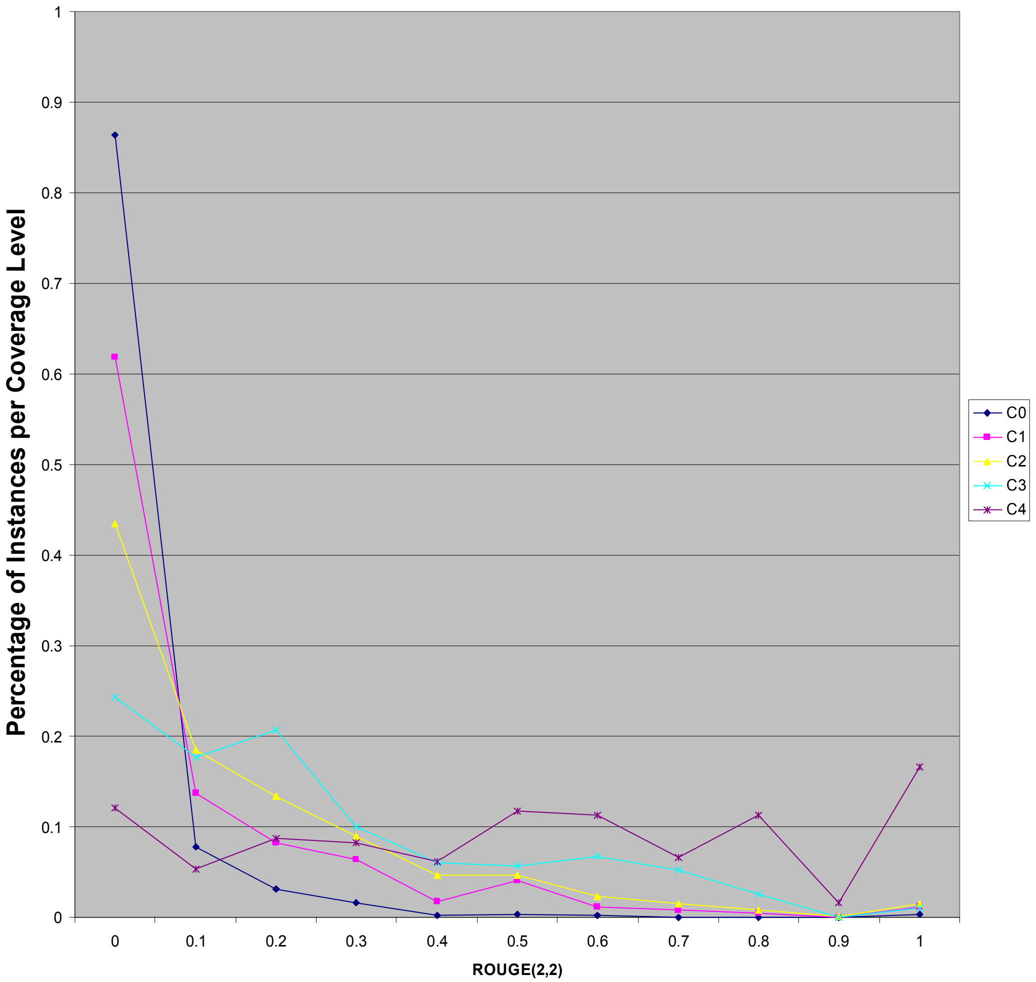
DUC 2001 - Single Doc Coverage vs. ROUGE(1,4) (SX) Distribution



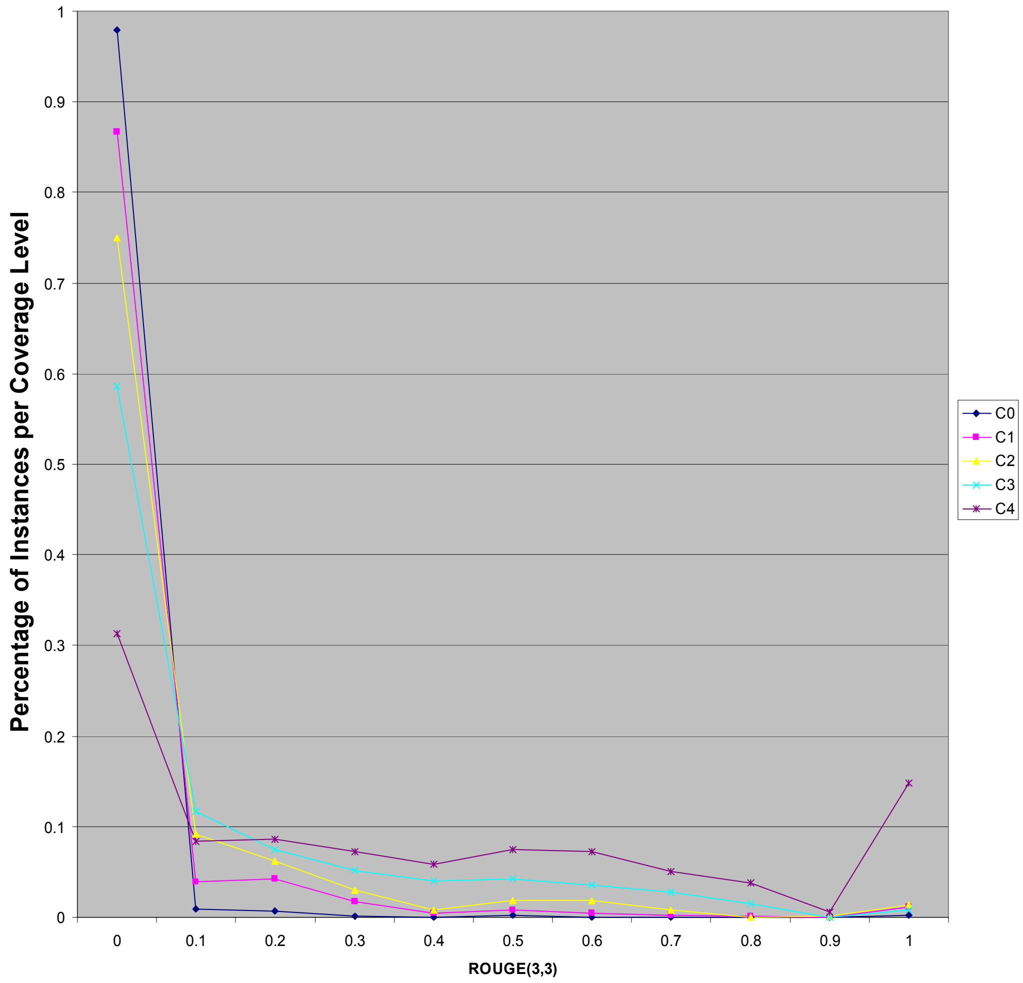
DUC 2001 - Single Doc Coverage vs. ROUGE(1,1) (SX) Distribution



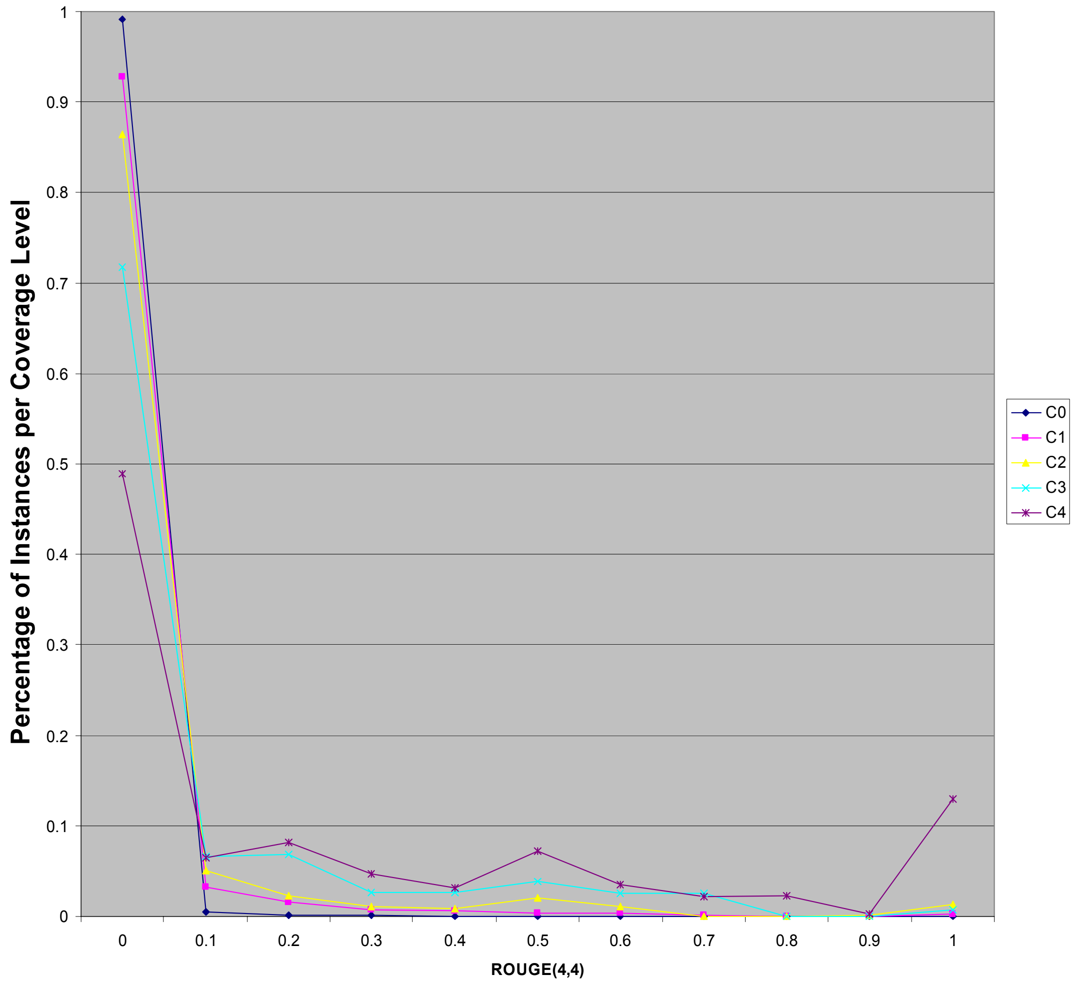
DUC 2001 - Single Doc Coverage vs. ROUGE(2,2) (SX) Distribution



DUC 2001 - Single Doc Coverage vs. ROUGE(3,3) (SX) Distribution



DUC 2001 - Single Doc Coverage vs. ROUGE(4,4) (SX) Distribution



DUC 2001 - ROUGE(1,4) Reliability of Significance Prediction

DUC01 Single-Doc without Stopwords - H0 Pairwise Observed Z-Score Coverage

	1	H	O	P	Q	R	S	T	V	W	X	Y	Z
1	0.000	1	-0.215	0.939	0.827	-0.545	2	-1.030	1.875	1.089	3	1.513	4
H		0.000	5	6	7	8	9	10	11	12	13	14	15
O			0.000	1.149	1.038	-0.324	16	-0.806	2.077	1.295	17	1.721	18
P				0.000	-0.112	-1.497	2.083	-1.984	0.954	0.166	2.022	0.570	2.297
Q					0.000	-1.383	2.194	-1.870	1.064	0.276	2.133	0.683	19
R						0.000	20	-0.490	21	1.640	22	2.081	23
S							0.000	24	-1.098	-1.880	-0.060	-1.529	0.237
T								0.000	25	2.119	26	27	28
V									0.000	-0.774	1.038	-0.399	1.319
W										0.000	1.820	0.393	2.093
X											0.000	-1.467	0.296
Y												0.000	1.750
Z													0.000

DUC01 Single-Doc without Stopwords - H0 Pairwise Observed Z-Score Ngram(1,4)

	1	H	O	P	Q	R	S	T	V	W	X	Y	Z
1	0.000		0.711	2.234	0.851	-0.068	1	-1.500	1.827	1.753	2	2.314	3
H		0.000		-1.701			-0.522		-2.189	-2.232	-0.869	-1.580	0.459
O			0.000	1.524	0.136	-0.777	4	-2.204	1.115	1.046	2.300	1.607	5
P				0.000	-1.395	-2.293	0.981		-0.412	-0.469	0.745	0.091	1.824
Q					0.000	-0.916			0.984	0.916	2.173	1.479	6
R						0.000	7	-1.427	1.887	1.814	8		9
S							0.000	10	-1.391	-1.439	-0.264	-0.885	0.816
T								0.000	11		12	13	14
V									0.000	-0.060	1.168	0.501	2.244
W										0.000	1.219	0.557	2.284
X											0.000	-0.648	1.115
Y												0.000	1.721
Z													0.000

$H_{Hs} = 28$

better

worse

Significant at α = .01

$H_{As} = 24$

$H_{hit} = 14$

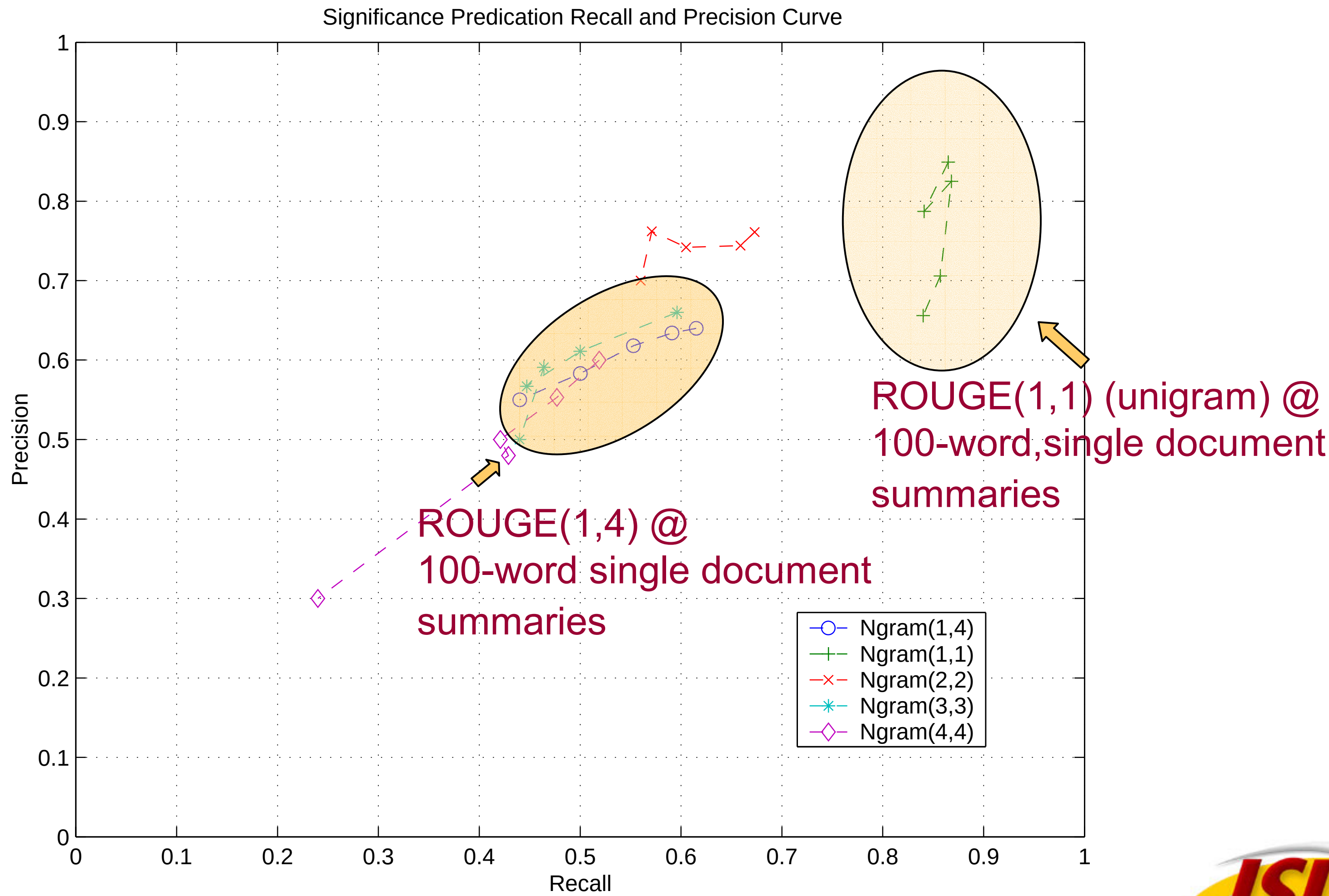
Recall $= H_{hit}/H_{Hs} = 0.500$

Precision $= H_{hit}/H_{As} = 0.583$

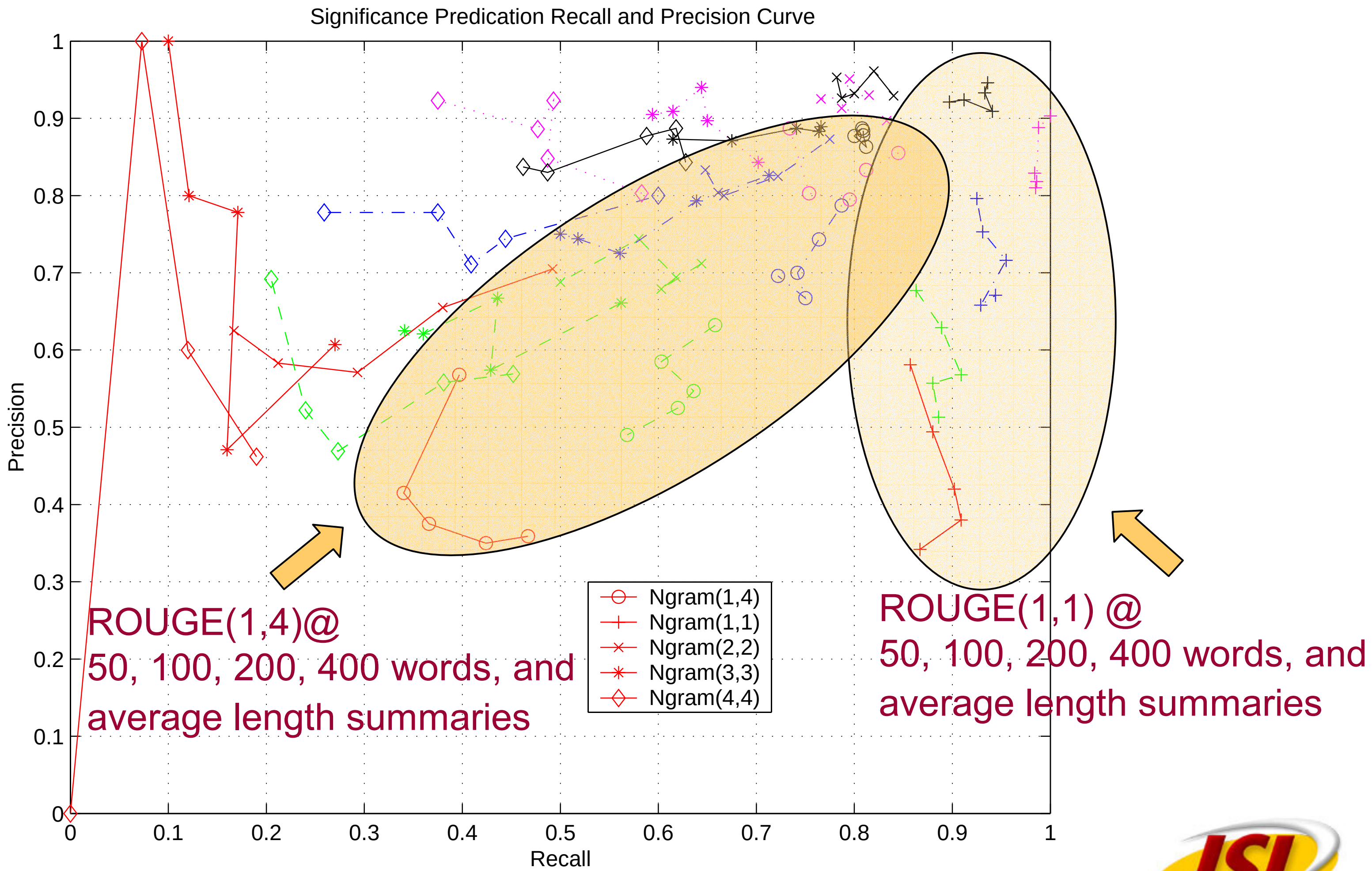
USC SCHOOL OF ENGINEERING

ISI Information Sciences Institute

DUC 2001 - ROUGE(i,j) Significance Predication (Single Document)



DUC 2001 - ROUGE(i,j) Significance Predication Multiple Documents



Summary of DUC 2003 Automatic Evaluation

- Correlation to the manual evaluation in Spearman ρ

ρ	Headline	Event
ROUGE(1,4)	0.725	0.623
ROUGE(1,1)	0.919	0.871
ROUGE(2,2)	0.801	0.868
ROUGE(3,3)	0.710	0.593
ROUGE(4,4)	0.623	0.281

- Reliability of significance prediction in recall, precision, and F1

	Headline	Event
ROUGE(1,4)	(0.581,0.742,0.652)	(0.229,0.917,0.366)
ROUGE(1,1)	(0.903,0.889,0.896)	(0.868,0.839,0.853)
ROUGE(2,2)	(0.750,0.861,0.802)	(0.694,0.990,0.816)
ROUGE(3,3)	(0.645,0.777,0.705)	(0.285,0.872,0.430)
ROUGE(4,4)	(0.556,0.734,0.633)	(0.201,0.935,0.331)

(Recall, Precision, F1)

DUC 2003 Headline Generation Task

- 13 systems with four different output formats
  (1) Sentences: S1, S9, S10, S15, S17, S22, S24
  (2) Clause, and phrases or words: S7, S13, S18, S25
  (3) Noun phrases: S21
  (4) Keywords: S26

Evaluation Results using only systems with sentence output

	ρ	Recall	Precision	F1
ROUGE(1,4)	1.000	0.800	1.000	0.889
ROUGE(1,1)	1.000	1.000	1.000	1.000
ROUGE(2,2)	1.000	0.900	1.000	0.947
ROUGE(3,3)	1.000	0.850	1.000	0.919
ROUGE(4,4)	1.000	0.750	1.000	0.857

Evaluation Results using only systems with sentence, clause plus Phrase or word output

	ρ	Recall	Precision	F1
ROUGE(1,4)	0.903	0.750	0.882	0.811
ROUGE(1,1)	0.939	0.975	0.929	0.951
ROUGE(2,2)	0.891	0.825	0.892	0.857
ROUGE(3,3)	0.903	0.800	0.842	0.820
ROUGE(4,4)	0.903	0.725	0.853	0.784

Evaluation Using Multiple Refs

- Perform Jackknifing experiments by leaving one out.
- Headline:
  - Best system vs. Human: 77%
- Event cluster:
  - Best system vs. Human: 84%

<u>Human vs. Human Task 1, Headline</u>
ROUGE(1,1) Average: 0.34299 (± 0.00994)
ROUGE(1,1) Median: 0.34981 (± 0.01021)
ROUGE(1,1) Maximum: 0.42459 (± 0.01075)
ROUGE(1,1) Minimum: 0.24777 (± 0.01017)

<u>System 17 vs. Human Task 1, Headline</u>
ROUGE(1,1) Average: 0.26554 (± 0.01393)
ROUGE(1,1) Median: 0.26581 (± 0.01384)
ROUGE(1,1) Maximum: 0.31461 (± 0.01586)
ROUGE(1,1) Minimum: 0.21591 (± 0.01262)

<u>Human vs. Human Task 2, Event Cluster</u>
ROUGE(1,1) Average: 0.34860 (± 0.02419)
ROUGE(1,1) Median: 0.34814 (± 0.02406)
ROUGE(1,1) Maximum: 0.39384 (± 0.02657)
ROUGE(1,1) Minimum: 0.30428 (± 0.02493)

<u>System 13 vs. Human Task 2, Event Cluster</u>
ROUGE(1,1) Average: 0.29317 (± 0.01817)
ROUGE(1,1) Median: 0.29365 (± 0.01748)
ROUGE(1,1) Maximum: 0.31274 (± 0.02110)
ROUGE(1,1) Minimum: 0.27264 (± 0.01803)

Conclusions & Future Directions

- ROUGE scores can be used to compare systems with similar output characteristics.
- ROUGE can be used for in-house hill-climbing for system development (Lin IRAL 2003)
- Next steps:
  - How to measure *quality* of summaries?
  - How to weight matches according to their importance?
  - How to accommodate the mismatch of words used in the human summaries and in the original documents?
  - How to adjust score according to summary length?
  - What is the effect of using different or multiple references?
  - Start collecting reference summaries.

Q&A

Thank You!