ROUGE: A Package for Automatic Evaluation of Summaries

Chin-Yew Lin

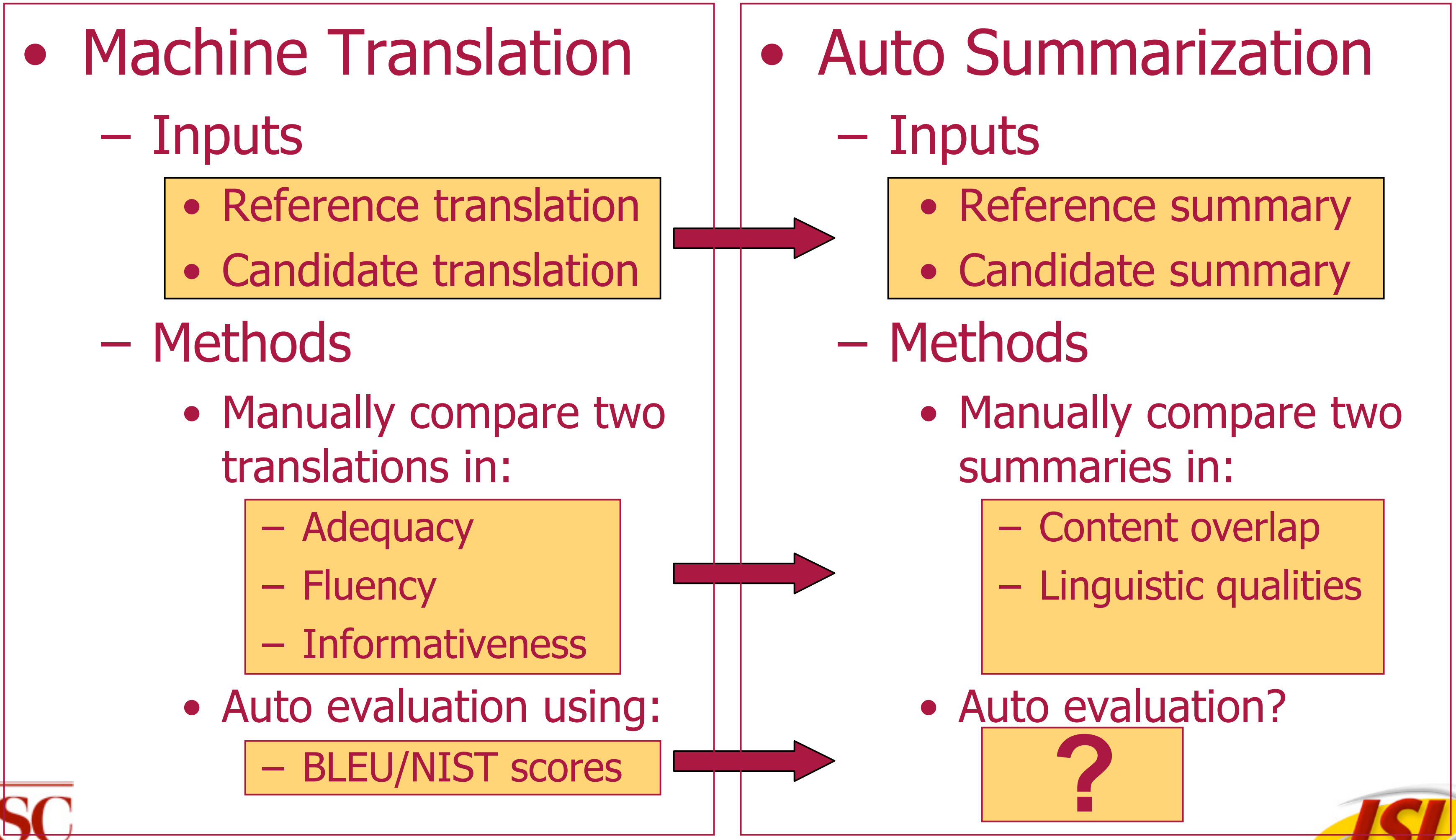
Information Sciences Institute

University of Southern California

Summarization Evaluation

- Basic assumptions
  - We know how to summarize.
  - We know what a good summary should be.
- The reality
  - Everyone summarizes.
  - Everyone has his/her own good summary.
- The question
  - Is objective evaluation of summarization possible, if everyone has his/her own good summary?

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

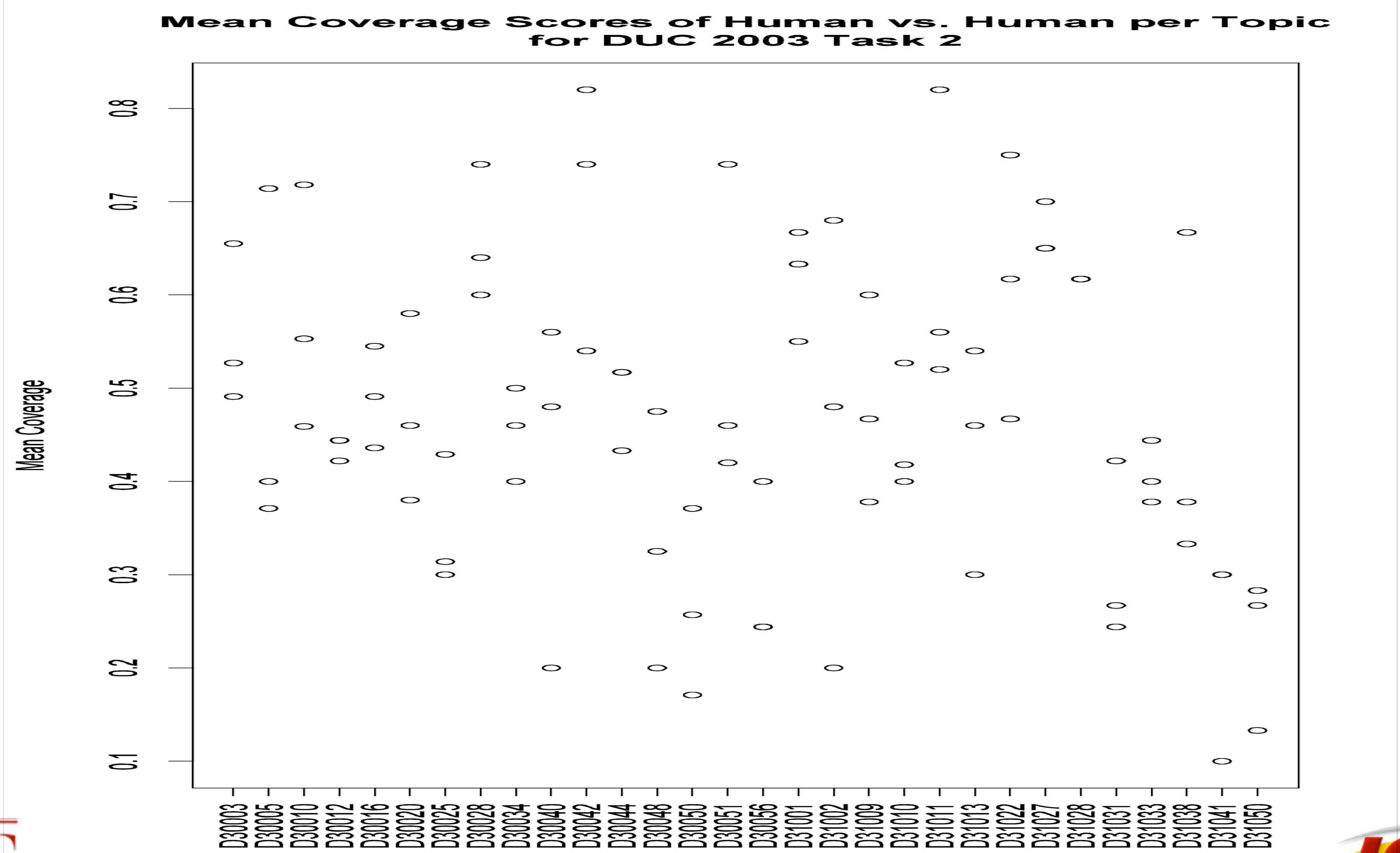
- Auto Summarization
  - Inputs
    - Reference summary
    - Candidate summary
  - Methods
    - Manually compare two summaries in:
      - Content overlap
      - Linguistic qualities
    - Auto evaluation?

?

Document Understanding Conference (DUC)

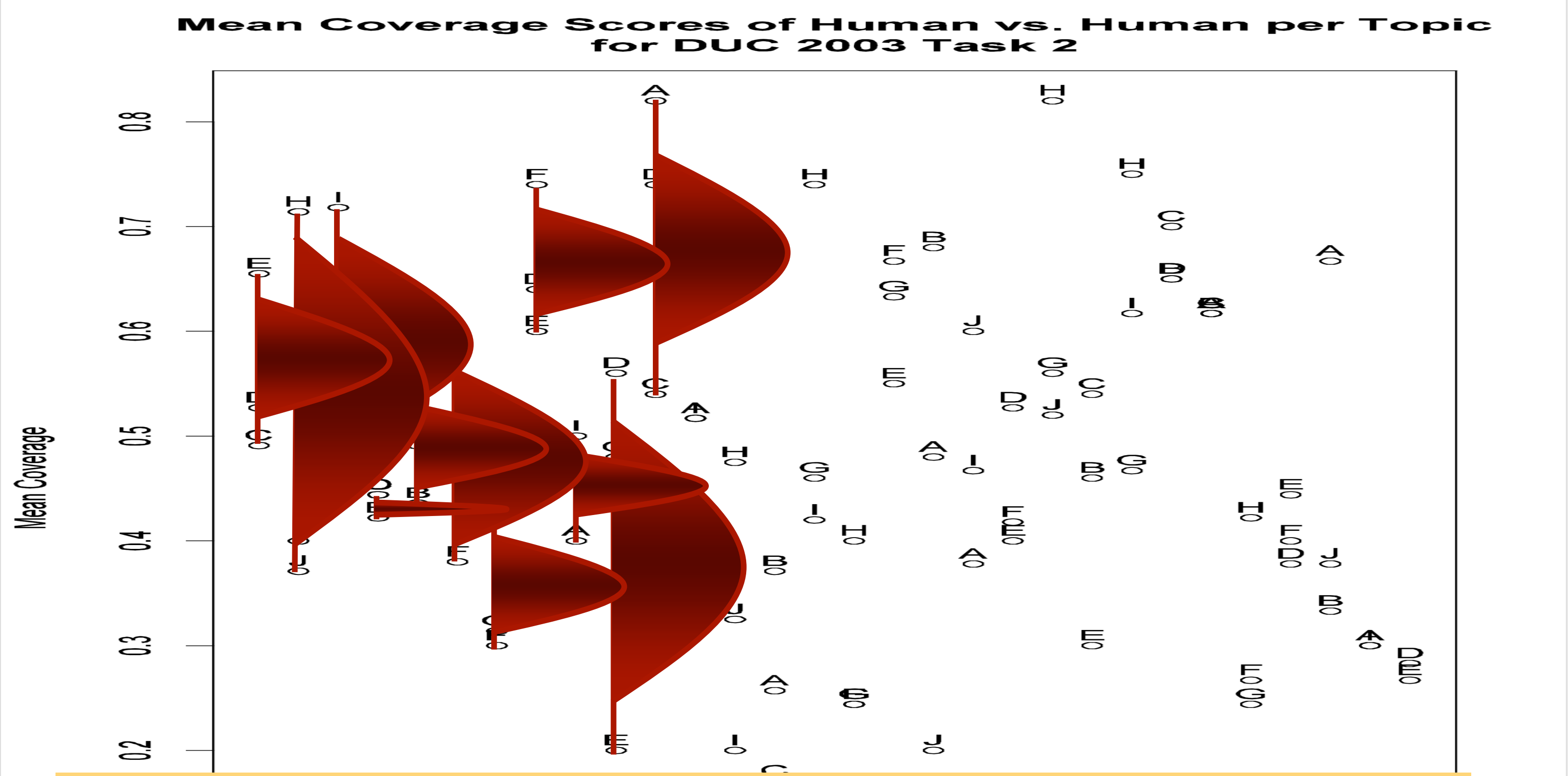
- Part of US DARPA TIDES Project DUC 01 - 04 (http://duc.nist.gov)
  - Tasks
    - Single-doc summarization (DUC 01 and 02: 30 topics)
    - Single-doc headline generation (DUC 03: 30 topics, 04: 50 topics)
    - Multi-doc summarization
      - Generic 10, 50, 100, 200 (2002) , and 400 (2001) words summaries
      - 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)
      - Short summaries of about 665 bytes in three different tasks in 2004
        - » focused by an event (50 TDT clusters)
        - » focused by an event but documents were translated into English from Arabic (24 topics)
        - » in response to a "who is X?" question (50 persons)
  - Participants
    - 15 systems in DUC 2001, 17 in DUC 2002, 21 in DUC 2003, and 25 in DUC 2004
- A new 3-year roadmap will be released during the summer.

DUC 2003 Human vs. Human (1)



Nenkova and Passonneau (HLT/NAACL 2004)

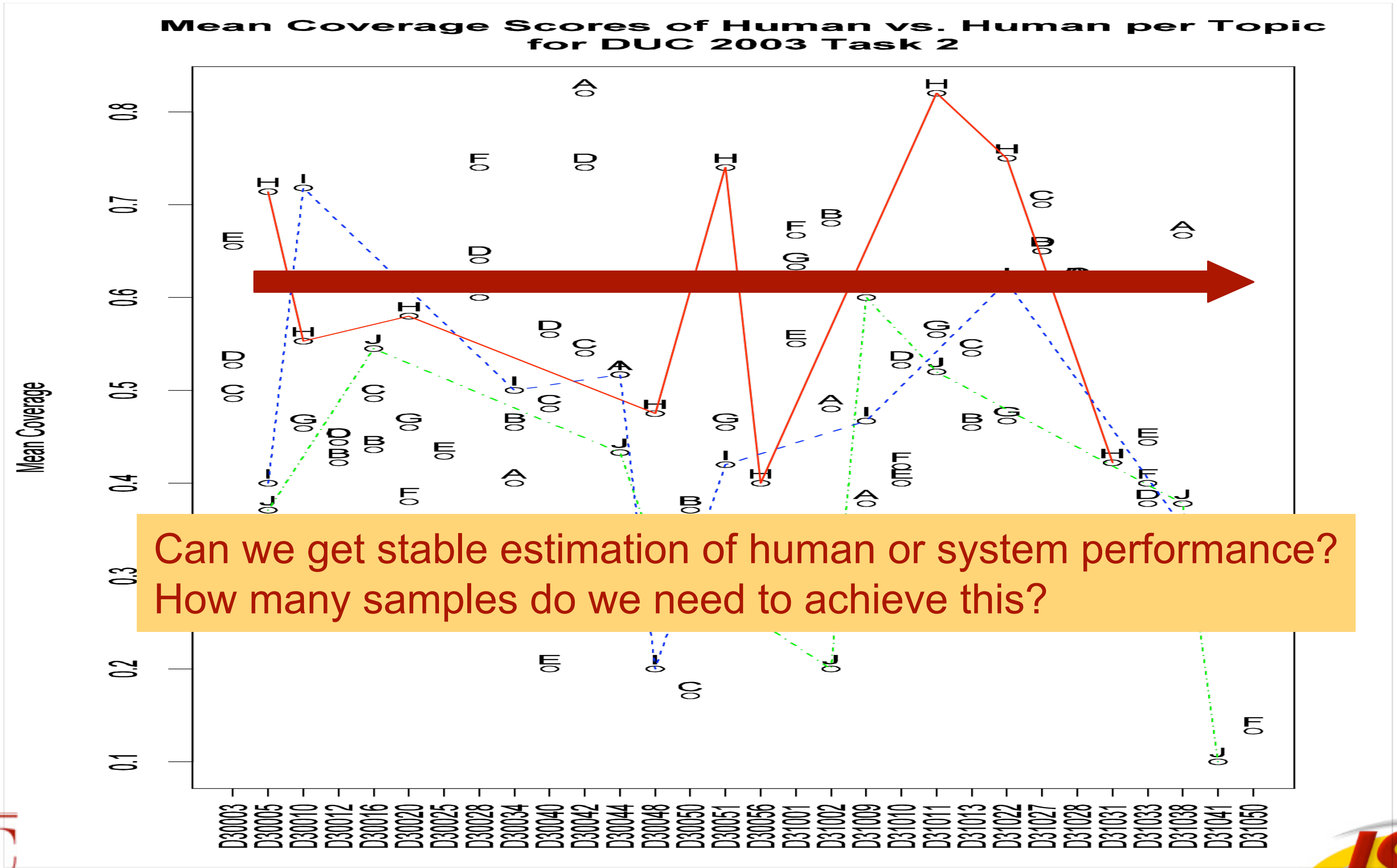
DUC 2003 Human vs. Human (2)



Mean Coverage Scores of Human vs. Human per Topic for DUC 2003 Task 2

1. Can we get consensus among humans?
2. If yes, how many humans do we need to get consensus?
3. Single reference or multiple references?

DUC 2003 Human vs. Human (3)



Can we get stable estimation of human or system performance?
How many samples do we need to achieve this?

Summary of Research Issues

- How to accommodate human inconsistency?
- Can we obtain stable evaluation results despite using only a single reference summary per evaluation?
- Will inclusion of multiple summaries make evaluation more or less stable?
- How can multiple references be used in improving stability of evaluations?
- How is stability of evaluation affected by sample size?

Recent Results

- Van Halteren and Teufel (2003)
  - Stable consensus factoid summary could be obtained if 40 to 50 reference summaries were considered.
    - 50 manual summaries of one text.
- Nenkova and Passonneau (2003)
  - Stable consensus semantic content unit (SCU) summary could be obtained if at least 5 reference summaries were used.
    - 10 manual multi-doc summaries for three DUC 2003 topics.
- Hori et al. (2003)
  - Using multiple references would improve evaluation stability if a metric taking into account consensus.
    - 50 utterances in Japanese TV broadcast news; each with 25 manual summaries.
- Lin and Hovy (2003), Lin (2004)
  - ROUGE, an automatic evaluation method used in summarization (DUC 2004) and MT (Lin and Och, ACL, COLING 2004).

Automatic Evaluation of Summarization Using ROUGE

- ROUGE summarization evaluation package
  - Currently (v1.4.2) include the following automatic evaluation methods:
    - ROUGE-N: N-gram based co-occurrence statistics
    - ROUGE-L: LCS-based statistics
    - ROUGE-W: Weighted LCS-based statistics that favors consecutive LCSes (see ROUGE note)
    - ROUGE-S: Skip-bigram-based co-occurrence statistics
    - ROUGE-SU: Skip-bigram plus unigram-based co-occurrence statistics
  - Free download for research purpose at: http://www.isi.edu/~cyl/ROUGE

ROUGE-N

- N-gram co-occurrences between reference and candidate translations.
  - Similar to BLEU in MT (Papineni et al. 2001)
- High order ROUGE-N with n-gram length greater than 1 estimates the fluency of summaries.
- Example:

  1. *police killed the gunman*
  2. police kill the gunman
  3. the gunman kill police

  ROUGE-N: S2=S3 (“police”, “the gunman”)

ROUGE-L

- Longest Common Subsequence (LCS)
  - Given two sequences X and Y, a longest common subsequence of X and Y is a common subsequence with maximum length.
  - Intuition
    - The longer the LCS of two translations is, the more similar the two translations are. (Saggion et al. 2002, MEAD)
  - Score
    - Use LCS-based recall score (ROUGE-L) to estimate the similarity between two translations. (see paper for more details)

ROUGE-L Example

- Example:
  1. *police killed the gunman*
  2. police kill the gunman
  3. the gunman kill police
- ROUGE-N: S2=S3 (“police”, “the gunman”)
- ROUGE-L:
  - S2=3/4 (“police the gunman”)
  - S3=2/4 (“the gunman”)
  - S2>S3

ROUGE-W

- Weighted Longest Common Subsequence
  - Example:
    - X: [A B C D E F G]
    - Y_1: [A B C D H I K]
    - Y_2: [A H B K C I D]
    - ROUGE-L(Y_1) = ROUGE-L(Y_2)
  - ROUGE-W favors strings with consecutive matches.
  - It can be computed efficiently using dynamic programming.

ROUGE-S

- Skip-Bigram
  - Any pair of words in their sentence order, allowing for arbitrary gaps.
  - Intuition
    - Consider long distance dependency.
    - Allow gaps in matches as LCS but count all in-sequence pairs; while LCS only counts the longest subsequences.
  - Score
    - Use skip-bigram-based recall score (ROUGE-S) to estimate the similarity between two translations. (see paper for more details)

ROUGE-S Example

- Example:
  1. *police killed the gunman*
  2. police kill the gunman
  3. the gunman kill police
  4. the gunman police killed
- ROUGE-N: S4>S2=S3
- ROUGE-L: S2>S3=S4
- ROUGE-S:
  - S2=3/6 (“police the”, “police gunman”, “the gunman”)
  - S3=1/6 (“the gunman”)
  - S4=2/6 (“the gunman”, “police killed”)
  - S2>S4>S3

Evaluation of ROUGE

- Corpora
  - DUC 01, 02, and 03 evaluation data
  - Including human and systems summaries
- Seven task formats
  - Single doc 10 and 100 words, multi-doc 10, 50, 100, 200, and 400 words
- Three versions
  - CASE: the original summaries
  - STEM: the stemmed version of summaries
  - STOP: STEM plus removal of stopwords
- Number of references
  - Single and different numbers of multiple references
- Quality criterion
  - Pearson's product moment correlation coefficients between systems' average ROUGE scores and their human assigned mean coverage score
- Metrics
  - 17 ROUGE metrics: ROUGE-N with N = 1 to 9, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU (with maximum skip-distance of 0, 4, and 9)
- Statistical significance
  - 95% confidence interval estimated using bootstrap resampling

100 Words Single-Doc Task

	DUC 2001 100 WORDS SINGLE DOC						DUC 2002 100 WORDS SINGLE DOC					
	1 REF			3 REFS			1 REF			2 REFS		
Method	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**
R-1	0.76	0.76	0.84	0.80	0.78	0.84	0.98	0.98	**0.99**	0.98	0.98	**0.99**
R-2	0.84	0.84	0.83	**0.87**	0.87	0.86	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-3	0.82	0.83	0.80	0.86	0.86	0.85	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-4	0.81	0.81	0.77	0.84	0.84	0.83	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**
R-5	0.79	0.79	0.75	0.83	0.83	0.81	**0.99**	**0.99**	0.98	**0.99**	**0.99**	0.98
R-6	0.76	0.77	0.71	0.81	0.81	0.79	0.98	**0.99**	0.97	**0.99**	**0.99**	0.98
R-7	0.73	0.74	0.65	0.79	0.80	0.76	0.98	0.98	0.97	**0.99**	**0.99**	0.97
R-8	0.69	0.71	0.61	0.78	0.78	0.72	0.98	0.98	0.96	**0.99**	**0.99**	0.97
R-9	0.65	0.67	0.59	0.76	0.76	0.69	0.97	0.97	0.95	0.98	0.98	0.96
R-L	0.83	0.83	0.83	0.86	0.86	0.86	**0.99**	**0.99**	0.99	**0.99**	**0.99**	**0.99**
R-S*	0.74	0.74	0.80	0.78	0.77	0.82	0.98	0.98	0.98	0.98	0.97	0.98
R-S4	0.84	**0.85**	0.84	**0.87**	**0.88**	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-S9	0.84	**0.85**	0.84	**0.87**	**0.88**	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-SU*	0.74	0.74	0.81	0.78	0.77	0.83	0.98	0.98	0.98	0.98	0.98	0.98
R-SU4	0.84	0.84	**0.85**	**0.87**	0.87	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-SU9	0.84	0.84	**0.85**	**0.87**	0.87	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-W-1.2	**0.85**	**0.85**	**0.85**	**0.87**	0.87	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**

10 Words Single-Doc Task

	DUC 2003 10 WORDS SINGLE DOC					
	1 REF	4REFS	1 REF	4 REFS	1 REF	4 REFS
Method	**CASE**		**STEM**		**STOP**	
R-1	0.96	0.95	0.95	0.95	0.90	0.90
R-2	0.75	0.76	0.75	0.75	0.76	0.77
R-3	0.71	0.70	0.70	0.68	0.73	0.70
R-4	0.64	0.65	0.62	0.63	0.69	0.66
R-5	0.62	0.64	0.60	0.63	0.63	0.60
R-6	0.57	0.62	0.55	0.61	0.46	0.54
R-7	0.56	0.56	0.58	0.60	0.46	0.44
R-8	0.55	0.53	0.54	0.55	0.00	0.24
R-9	0.51	0.47	0.51	0.49	0.00	0.14
R-L	**0.97**	**0.96**	**0.97**	**0.96**	0.97	0.96
R-S*	0.89	0.87	0.88	0.85	0.95	0.92
R-S4	0.88	0.89	0.88	0.88	0.95	0.96
R-S9	0.92	0.92	0.92	0.91	0.97	0.95
R-SU*	0.93	0.90	0.91	0.89	0.96	0.94
R-SU4	**0.97**	**0.96**	0.96	0.95	**0.98**	**0.97**
R-SU9	**0.97**	0.95	0.96	0.94	0.97	0.95
R-W-1.2	0.96	**0.96**	0.96	**0.96**	0.96	0.96

100 Words Multi-Doc Task

	(A1) DUC 2001 100 WORDS MULTI						(A2) DUC 2002 100 WORDS MULTI						(A3) DUC 2003 100 WORDS MULTI					
	1 RFF			3 REFS			1 REF			2 REFS			1 REF			4 REFS		
Method	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP
R-1	0.48	0.56	**0.86**	0.53	0.57	**0.87**	0.66	0.66	0.77	0.71	0.71	0.78	0.58	0.57	0.71	0.58	0.57	0.71
R-2	0.55	**0.57**	0.64	0.59	**0.61**	0.71	0.83	0.83	**0.80**	0.88	0.87	**0.85**	**0.69**	**0.67**	0.71	**0.79**	**0.79**	**0.81**
R-3	0.46	0.45	0.47	0.53	0.53	0.55	**0.85**	**0.84**	0.76	**0.89**	**0.88**	0.83	0.54	0.51	0.48	0.76	0.75	0.74
R-4	0.39	0.39	0.43	0.48	0.49	0.47	0.80	0.80	0.63	0.83	0.82	0.75	0.37	0.36	0.36	0.62	0.61	0.52
R-5	0.38	0.39	0.33	0.47	0.48	0.43	0.73	0.73	0.45	0.73	0.73	0.62	0.25	0.25	0.27	0.45	0.44	0.38
R-6	0.39	0.39	0.20	0.45	0.46	0.39	0.71	0.72	0.38	0.66	0.64	0.46	0.21	0.21	0.26	0.34	0.31	0.29
R-7	0.31	0.31	0.17	0.44	0.44	0.36	0.63	0.65	0.33	0.56	0.53	0.44	0.20	0.20	0.23	0.29	0.27	0.25
R-8	0.18	0.19	0.09	0.40	0.40	0.31	0.55	0.55	0.52	0.50	0.46	0.52	0.18	0.18	0.21	0.23	0.22	0.23
R-9	0.11	0.12	0.06	0.38	0.38	0.28	0.54	0.54	0.52	0.45	0.42	0.52	0.16	0.16	0.19	0.21	0.21	0.21
R-L	**0.49**	0.49	0.49	0.56	0.56	0.56	0.62	0.62	0.62	0.65	0.65	0.65	0.50	0.50	0.50	0.53	0.53	0.53
R-S*	0.45	0.52	0.84	0.51	0.54	0.86	0.69	0.69	0.77	0.73	0.73	0.79	0.60	0.60	0.67	0.61	0.60	0.70
R-S4	0.46	0.50	0.71	0.54	0.57	0.78	0.79	0.80	0.79	0.84	0.85	0.82	0.63	0.64	0.70	0.73	0.73	0.78
R-S9	0.42	0.49	0.77	0.53	0.56	0.81	0.79	0.80	0.78	0.83	0.84	0.81	0.65	0.65	0.70	0.70	0.70	0.76
R-SU*	0.45	0.52	0.84	0.51	0.54	**0.87**	0.69	0.69	0.77	0.73	0.73	0.79	0.60	0.59	0.67	0.60	0.60	0.70
R-SU4	0.47	0.53	0.80	0.55	0.58	0.83	0.76	0.76	0.79	0.80	0.81	0.81	0.64	0.64	**0.74**	0.68	0.68	0.76
R-SU9	0.44	0.50	0.80	0.53	0.57	0.84	0.77	0.78	0.78	0.81	0.82	0.81	0.65	0.65	0.72	0.68	0.68	0.75
R-W-1.2	0.52	0.52	0.52	**0.60**	0.60	0.60	0.67	0.67	0.67	0.69	0.69	0.69	0.53	0.53	0.53	0.58	0.58	0.58

Multi-Doc Task of Different Summary Sizes

	(C) DUC02 10			(D1) DUC01 50			(D2) DUC02 50			(E1) DUC01 200			(E2) DUC02 200			(F) DUC01 400		
Method	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**	**CASE**	**STEM**	**STOP**
R-1	0.71	0.68	0.49	0.49	0.49	0.73	0.44	0.48	**0.80**	0.81	0.81	**0.90**	0.84	0.84	0.91	0.74	0.73	**0.90**
R-2	0.82	0.85	**0.80**	0.43	0.45	0.59	0.47	0.49	0.62	**0.84**	**0.85**	0.86	**0.93**	**0.93**	**0.94**	**0.88**	**0.88**	0.87
R-3	0.59	0.74	0.75	0.32	0.33	0.39	0.36	0.36	0.45	0.80	0.80	0.81	0.90	0.91	0.91	0.84	0.84	0.82
R-4	0.25	0.36	0.16	0.28	0.26	0.36	0.28	0.28	0.39	0.77	0.78	0.78	0.87	0.88	0.88	0.80	0.80	0.75
R-5	-0.25	-0.25	-0.24	0.30	0.29	0.31	0.28	0.30	0.49	0.77	0.76	0.72	0.82	0.83	0.84	0.77	0.77	0.70
R-6	0.00	0.00	0.00	0.22	0.23	0.41	0.18	0.21	-0.17	0.75	0.75	0.67	0.78	0.79	0.77	0.74	0.74	0.63
R-7	0.00	0.00	0.00	0.26	0.23	0.50	0.11	0.16	0.00	0.72	0.72	0.62	0.72	0.73	0.74	0.70	0.70	0.58
R-8	0.00	0.00	0.00	0.32	0.32	0.34	-0.11	-0.11	0.00	0.68	0.68	0.54	0.71	0.71	0.70	0.66	0.66	0.52
R-9	0.00	0.00	0.00	0.30	0.30	0.34	-0.14	-0.14	0.00	0.64	0.64	0.48	0.70	0.69	0.59	0.63	0.62	0.46
R-L	0.78	0.78	0.78	**0.56**	**0.56**	0.56	0.50	0.50	0.50	0.81	0.81	0.81	0.88	0.88	0.88	0.82	0.82	0.82
R-S*	0.83	0.82	0.69	0.46	0.45	**0.74**	0.46	0.49	**0.80**	0.80	0.80	**0.90**	0.84	0.85	0.93	0.75	0.74	0.89
R-S4	**0.85**	**0.86**	0.76	0.40	0.41	0.69	0.42	0.44	0.73	0.82	0.82	0.87	0.91	0.91	0.93	0.85	0.85	0.85
R-S9	0.82	0.81	0.69	0.42	0.41	0.72	0.40	0.43	0.78	0.81	0.82	0.86	0.90	0.90	0.92	0.83	0.83	0.84
R-SU*	0.75	0.74	0.56	0.46	0.46	**0.74**	0.46	0.49	**0.80**	0.80	0.80	**0.90**	0.84	0.85	0.93	0.75	0.74	0.89
R-SU4	0.76	0.75	0.58	0.45	0.45	0.72	0.44	0.46	0.78	0.82	0.83	0.89	0.90	0.90	0.93	0.84	0.84	0.88
R-SU9	0.74	0.73	0.56	0.44	0.44	0.73	0.41	0.45	0.79	0.82	0.82	0.88	0.89	0.89	0.92	0.83	0.82	0.87
R-W-1.2	0.78	0.78	0.78	**0.56**	**0.56**	0.56	**0.51**	**0.51**	0.51	**0.84**	0.84	0.84	0.90	0.90	0.90	0.86	0.86	0.86

Summary of Results

- Overall
  - Using multiple references achieved better correlation with human judgment than just using a single reference.
  - Using more samples achieved better correlation with human judgment (DUC 02 vs. other DUC data).
  - Stemming and removing stopwords improved correlation with human judgment.
  - Single-doc task had better correlation than multi-doc
- Specific
  - ROUGE-S4, S9, and ROUGE-W1.2 were the best in 100 words single-doc task, but were statistically indistinguishable from most other ROUGE metrics.
  - ROUGE-1, ROUGE-L, ROUGE-SU4, ROUGE-SU9, and ROUGE-W1.2 worked very well in 10 words headline like task (Pearson's ρ ~ 97%).
  - ROUGE-1, 2, and ROUGE-SU* were the best in 100 words multi-doc task but were statistically equivalent to other ROUGE-S and SU metrics.
  - ROUGE-1, 2, ROUGE-S, and SU worked well in other multi-doc tasks.

Ongoing Work

- Summary and sentence level error analysis
  - Summary level
    - Evaluate techniques used in ETS' E-Rater and its successors in automatic evaluation of summaries.
  - Sentence level
    - Matching at concept level instead of lexical level:
      - Synonyms and paraphrases
      - Utilize consensus in reference summaries
    - Matching at syntactic level
      - Dependency structure based co-occurrence statistics
- Large scale reference summary corpus creation

Q&A

Thank You!