Content Realization: Linguistic Quality

Ling 573
Systems and Applications
May 7, 2018

UNIVERSITY OF WASHINGTON

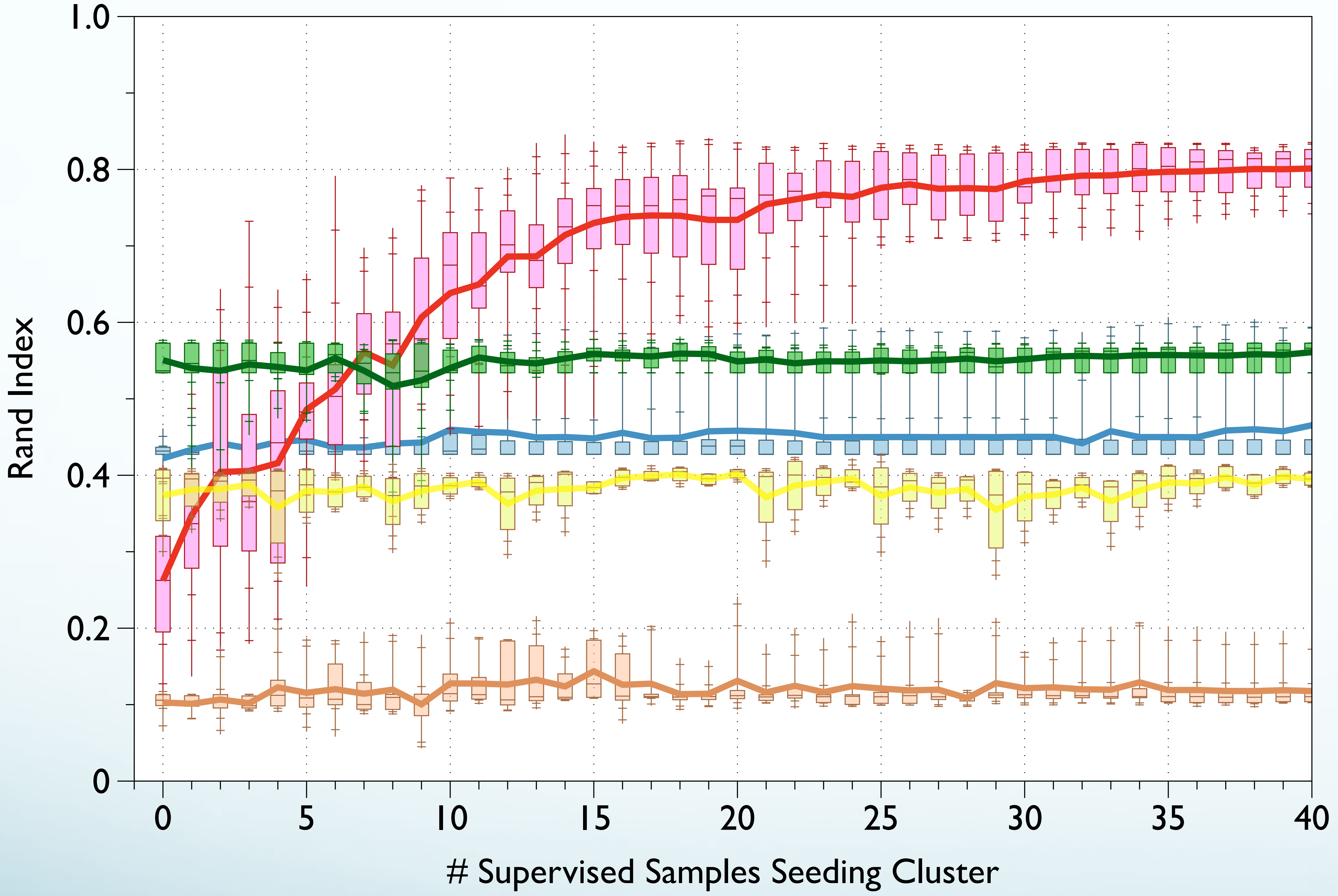
PROFESSIONAL MASTER'S IN
COMPUTATIONAL LINGUISTICS

Begin Recording

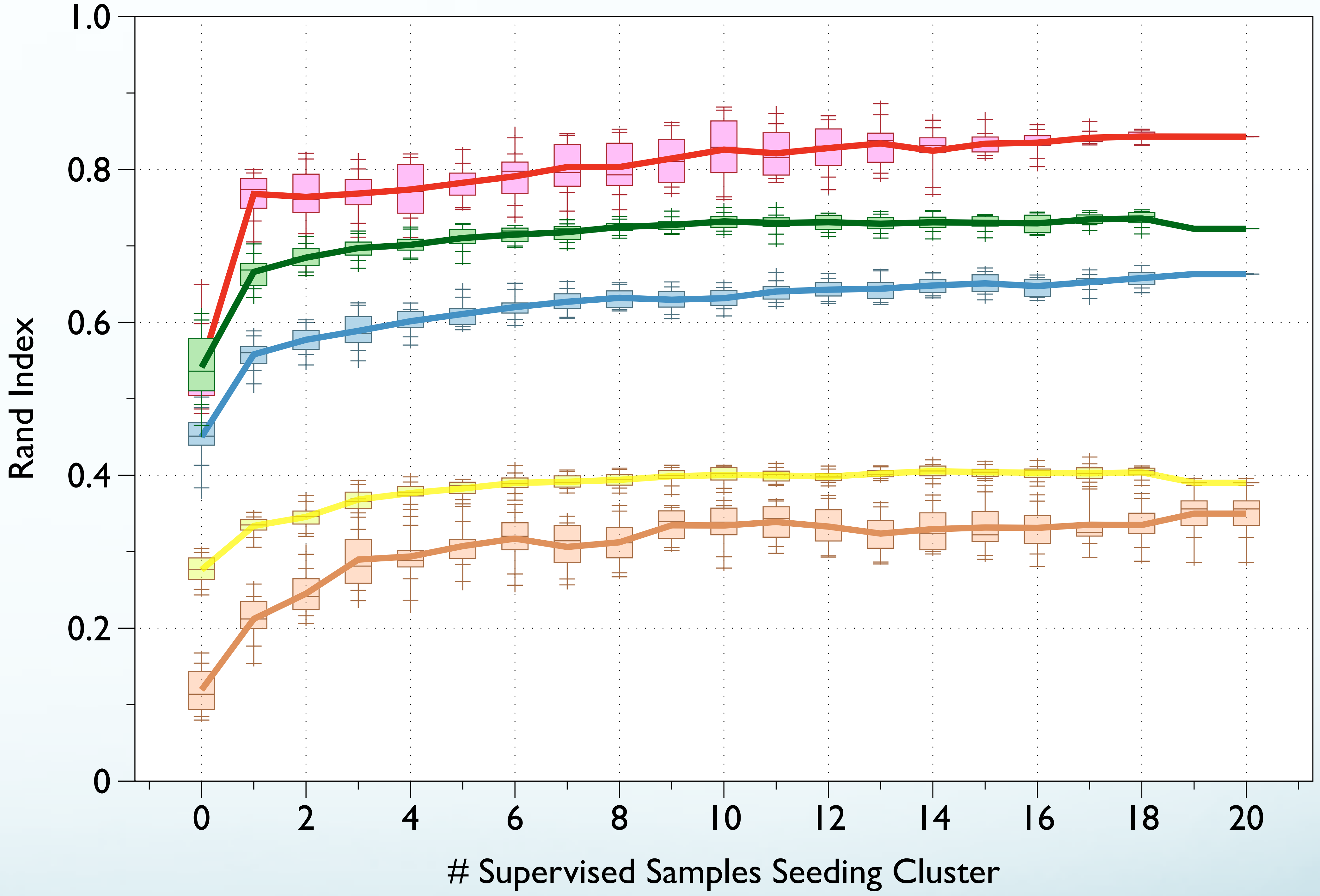
Miscellanea

Semi-Supervised Clustering Revisited

KMeans Clustering on TAC Categories
5 Clusters, 20 Runs



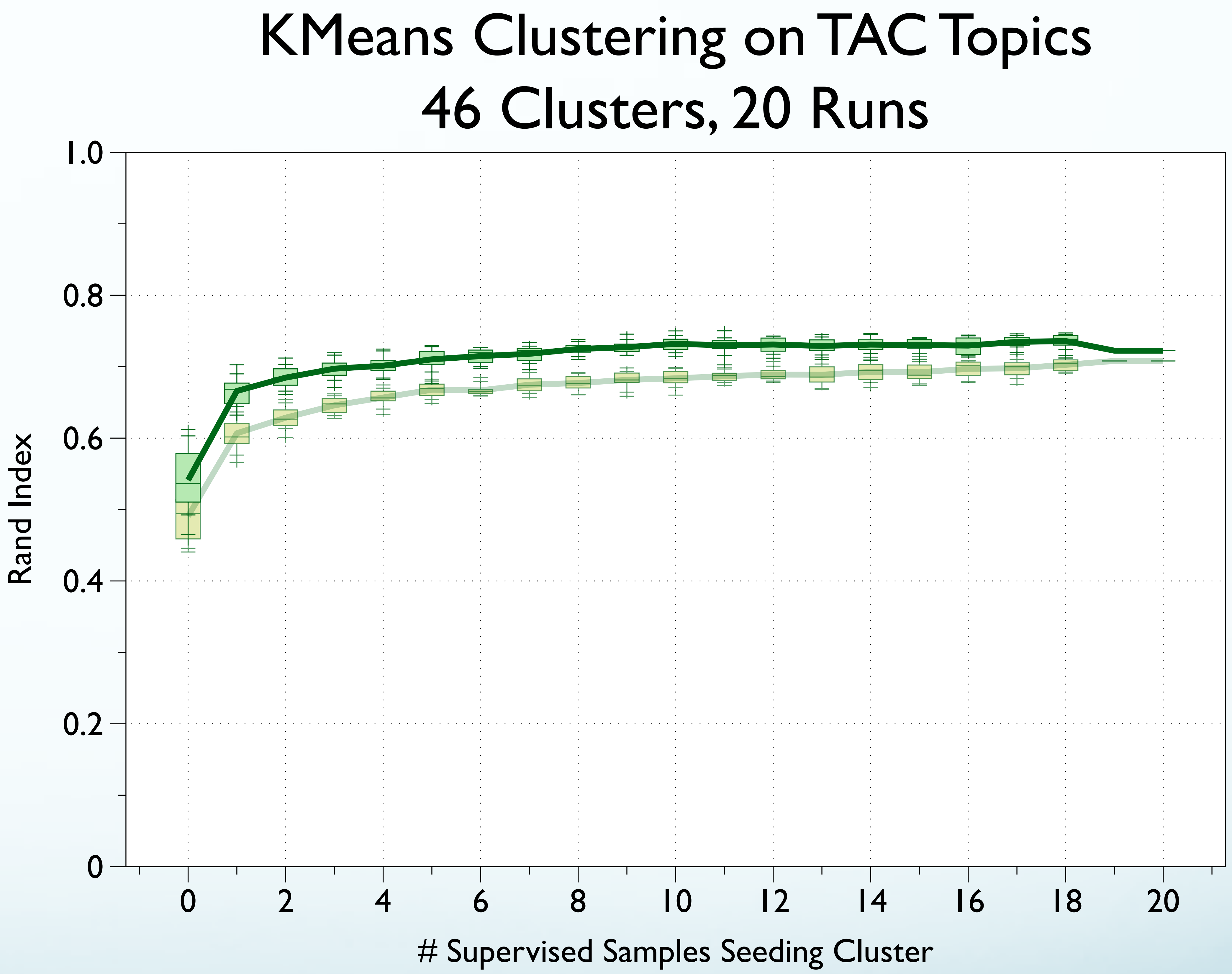
KMeans Clustering on TAC Topics
46 Clusters, 20 Runs



= tf•idf via sklearn
= GloVe via spacy
= Google News via gensim (100B words)
= Doc2Vec on ENG-GW (~300M)
= Word2Vec on ENG-GW (~30B)

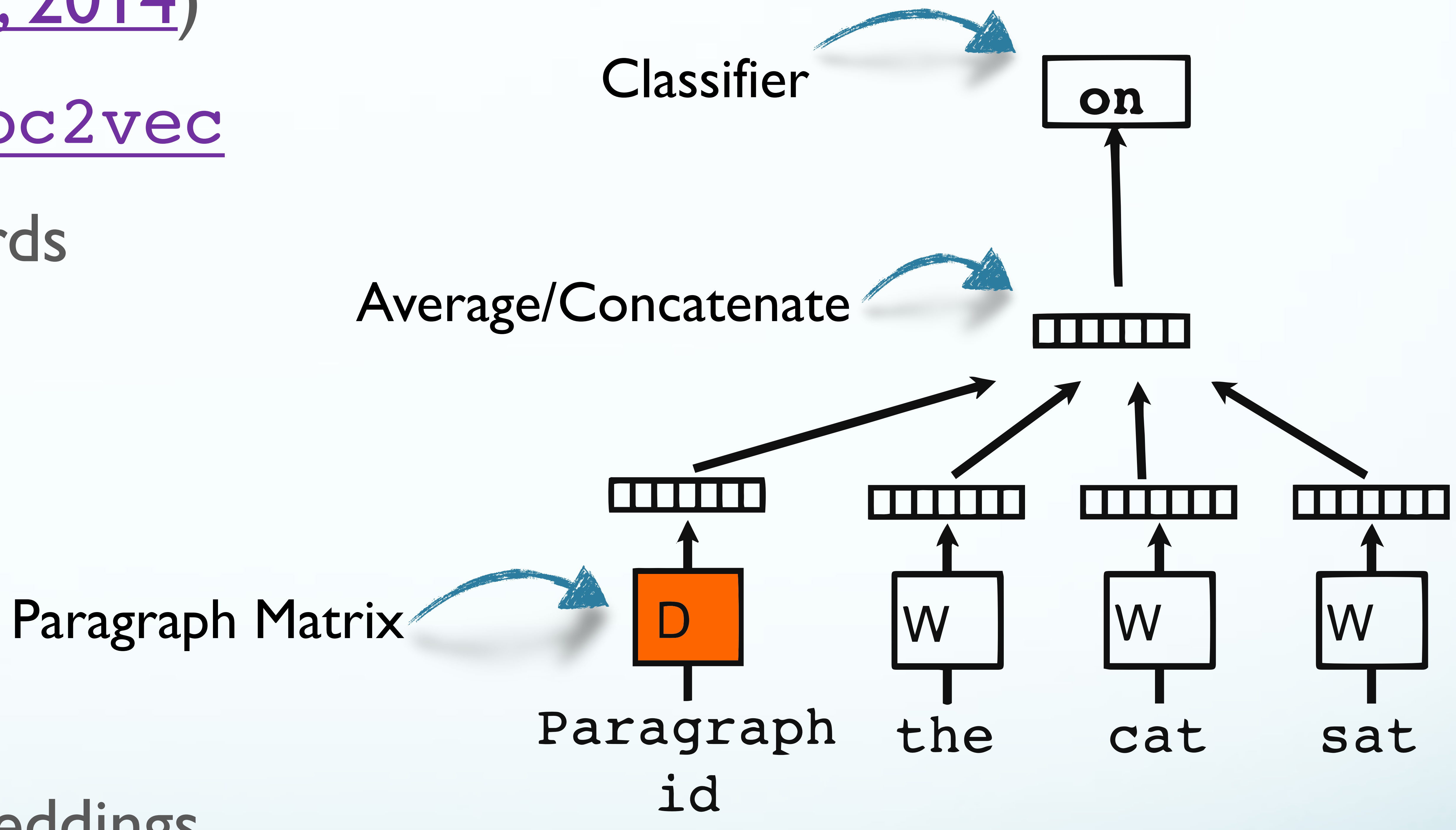
Semi-Supervised Clustering Revisited

- Lower Line:
  - lowercasing with Google News embeddings
  - …they are not lowercased!



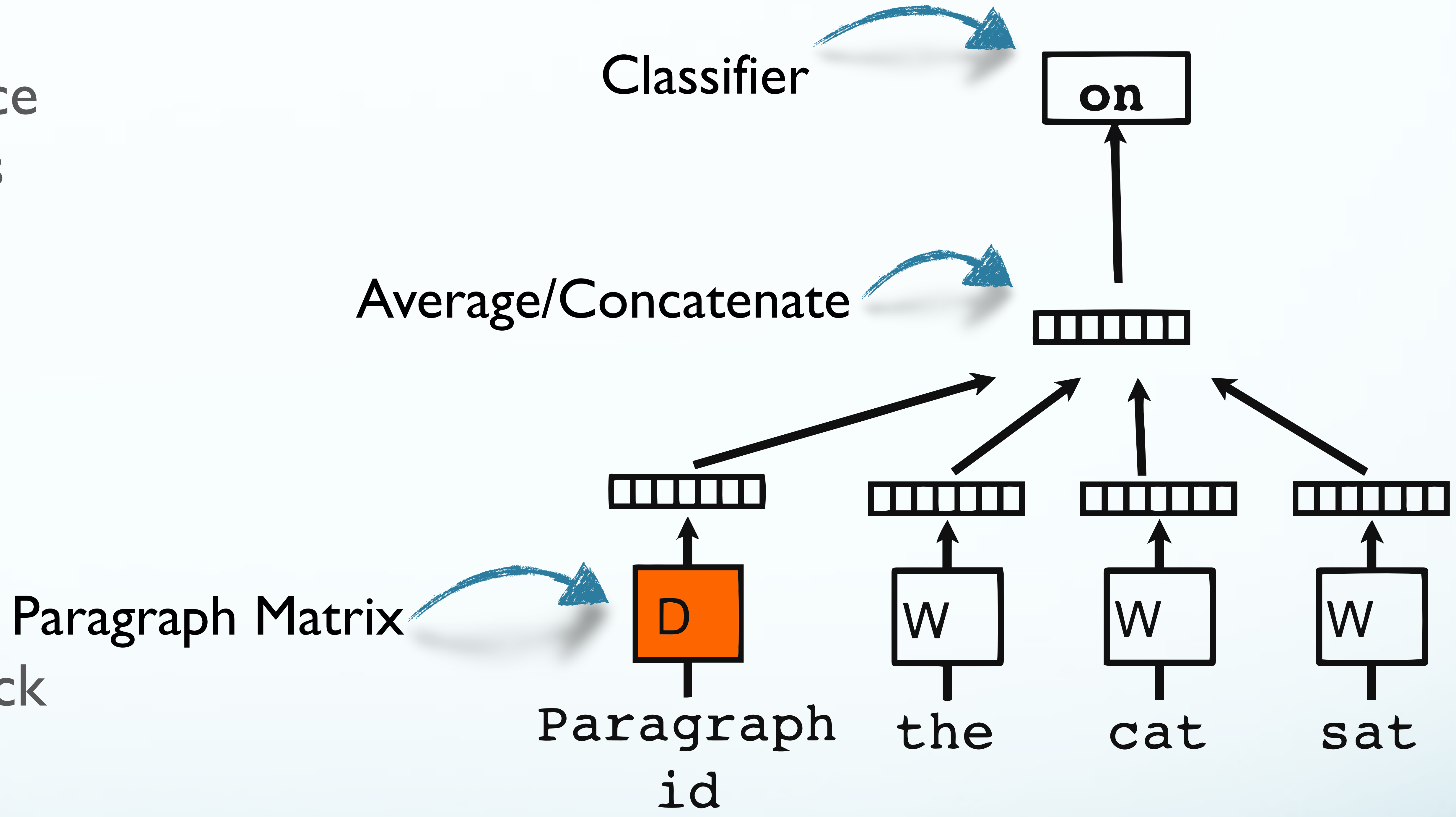
Other Embeddings to Explore

- doc2vec / paragraph2vec (Le & Mikolov, 2014)
  - Implemented in **`gensim`** as `models.doc2vec`
  - Add arbitrary, unique tag to group of words
    - Sentence ID
    - Paragraph ID
    - Document ID
- Main Idea:
  - Associating unique ID as part of the embeddings pressures network to learn that groups have topicality



Other Embeddings to Explore

- In theory:
  - These embeddings could help represent sentence similarity better than mean of word embeddings
  - Use for:
    - Content selection / Redundancy Management
    - Topic Clustering / Coherence
- In practice…
  - Seems to suffer from the Machine Learning "Black Box" problem
  - Difficulties in reproducibility
    - Artificial intelligence faces reproducibility crisis, Hutson (2018)



Other Resources

- I have placed Google news Word2Vec pre-trained vectors in:
  - `/dropbox/17-18/573/other_resources/word_embeddings GoogleNews-vectors-negative300.bin.gz`
  - Load with `gensim.models.KeyedVectors.load($PATH, binary=`**`True`**`)`
- You MUST use condor!
  - This model will take **~3.4GB** when loaded into memory.

Other Resources

- gensim Word2Vec embeddings trained on ENG_GW (superset of AQUAINT-2)
  - `/dropbox/17-18/573/other_resources/word_embeddings/eng_gw eng_gw/eng_gw_lower`
  - load with `gensim.models.Word2Vec.load($PATH)`
  - Lowercased (non-cased still running)

Other Resources

- Newsroom corpus (Grusky et al, 2018) (Webpage: summari.es)
  - Giant crawl of news stories
  - Use HTML <meta> tags written by humans as single-document summaries
  - `/dropbox/17-18/573/other_resources/newsroom`
    - (As it becomes available)

Roadmap

- Content Realization in Summarization
  - Goals
  - Broad Approaches
- Readability and Linguistic Quality
  - Corpus study and analysis
  - Automatic Evaluation
  - Improvements for MDS

Goals of Content Realization:
Abstractive Summaries

- Content selection works over concepts
- Need to produce important concepts in fluent NL

Goals of Content Realization:
Extractive Summaries

- Draw from existing NL sentences
- Extreme compression — e.g. 60 byte summaries (headlines)
- Maximize ***density*** of ***relevant*** information
  - Remove verbose, unnecessary, or redundant content
- Increase readability, fluency, linguistic quality
  - Present content from multiple docs, non-adjacent sentences

Broad Approaches:
Abstractive Summaries

- **Complex Q-A**
  - Template-based methods
- **More Desirable**
  - Full Natural Lnguage Generation (NLG)
  - Concept-to-text

Broad Approaches:
Extractive Summaries

- **Sentence compression**
  - Remove “unnecessary” phrases within sentences
- **Sentence reformulation**
  - Modify portions of sentence for coreference, readability, etc
- **Sentence fusion**
  - Merge content from multiple sentences

Linguistic Quality

Shared Tasks

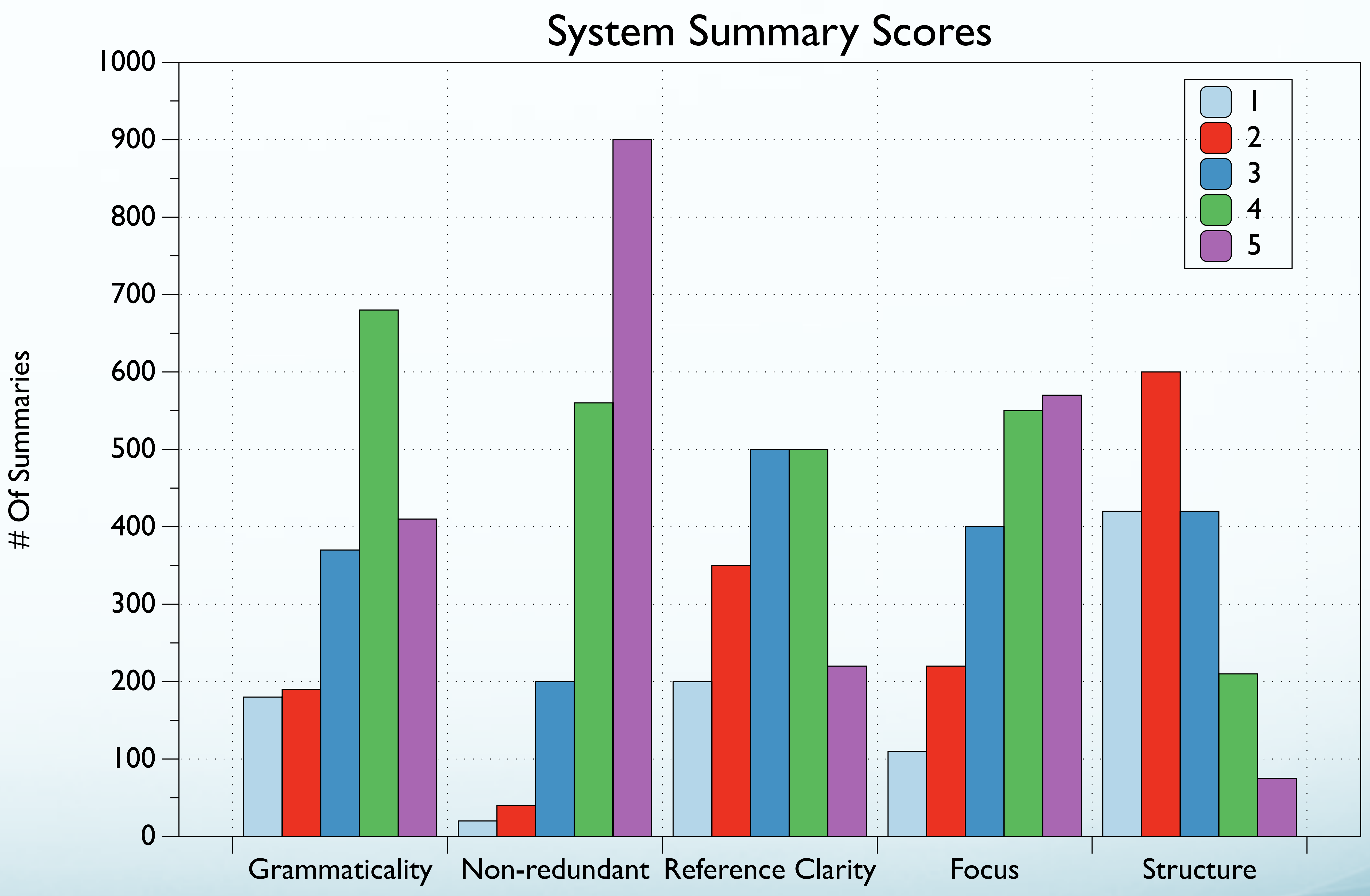
- Take content as primary evaluation measure
  - ROUGE, Pyramid, Responsiveness
  - Linguistic quality also part of formal evaluation
- Tac “readability”
  - Scored manually on five-point Likert scale
  - Aims to capture readability, fluency
    - Independent of summary content

What is “Readability?”

- **Grammaticality**:
  - No fragments, datelines, ill-formed sentences, etc.
- **Non-redundancy**
  - No unnecessary repetition: includes content, sentences, or full NPs when pronoun is better
- **Referential clarity**
  - Both presence/salience of antecedents, relevance of items
- **Focus**
  - Only content related to summary
- **Coherence**: “Well-structured”

Score Distributions

- DUC 2006 results:



What is "Readability"

- Definition subsumes many phenomena
- What types of errors do these systems make?
- What errors, issues, are reflected in the scores?
- LQVSumm (Friedrich et al, 2014)
  - Annotate linguistic "violations" in automatic summaries
    - TAC2011 data: ~2000 peer summaries
    - Categorize and tabulate
  - Assess correlation with Readability scores

Example

- “the girls” + “the Amish School”
  - Missing entity introduction
- “Miller” said
  - Needs explanation
- “The gunman, a local truck driver…”
  - Already introduced

Charles Carl Roberts IV may have planned to molest the girls at the Amish school, but police have no evidence that he actually did. Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. The suspect apparently called his wife from a cell phone shortly before the shooting began, saying ***he was “acting out in revenge for something that happened 20 years ago****, Miller said. The gunman, a local truck driver Charles Roberts, was apparently* ***acting in “revenge” for an incident that happened to him 20 years ago.***

Friedrich et al, 2014, p. 1591, Fig. 1

Violation Categories

- Entity Mentions
  - Affect coreference and readability

FM_EXPL	First mention without explanation
SM+EXPL	Subsequent mention with explanation
DNP_REF	Definite NP without previous mention
INP+REF	Indefinite NP with previous mention
PRN+MISSA	Pronoun with missing antecedent
PRN+MISSLA	Pronoun with misleading antecedent
ACR_EXPL	Acronym without explanation

Violation Categories

- Clausal level
  - Arbitrary spans — up to sentence level

INCOMPLSN	Incomplete sentence
INCLDATE	Included dateline info
OTHRUNGR	Other ungrammatical
NOSEMREL	No semantic relation between sentences
NODISREL	Discourse relation doesn't fit
REDUNINF	Redundant information

corpus	G-Flow 50 documents		TAC 1,935 document				
violation type	count	avg/doc	count	avg/doc	**Pearson's *r*** Readability	Pyramid	Respons.
entity level violations							
DNP-REF	3	0.06	958	0.50	**-0.122**	**-0.166**	**-0.133**
FM-EXPL	6	0.12	792	0.41	0.006	-0.050	**-0.066**
INP+REF	1	0.02	430	0.22	-0.052	**0.235**	**0.109**
PRN+MISSA	2	0.04	361	0.19	**-0.191**	**-0.140**	**-0.156**
SM+EXPL	1	0.02	162	0.08	0.020	**0.089**	0.045
PRN+MISLA	0	0.00	27	0.01	**-0.065**	**-0.073**	**-0.089**
ACR-EXPL	3	0.04	11	0.01	-0.038	-0.056	-0.006
sum(DNP-REF, PRN+MISSA)	5	0.1	1319	0.68	**-0.204**	**-0.208**	**-0.192**
sum(entity level violations)	15	0.03	2741	1.42	**-0.167**	-0.074	**-0.127**
clause level violations							
INCOMPLSN	0	0.00	1,044	0.54	**-0.210**	0.000	-0.029
OTHRUNGR	3	0.06	793	0.41	**-0.180**	0.007	-0.016
INCLDATE	3	0.06	412	0.21	**-0.090**	0.039	0.051
REDUNDINF	3	0.06	504	0.26	**-0.160**	**0.156**	**0.077**
NOSEMREL	0	0.00	142	0.07	**-0.148**	**-0.102**	**-0.132**
NODISREL	1	0.02	91	0.05	-0.025	**-0.081**	**-0.062**
misleading discourse connectives⋆	1	0.02	114	0.06	-	-	-
sum(clause level violations)	10	0.2	2,986	1.54	**-0.325**	0.041	-0.016
sum(clause level violations, DNP-REF, PRN+MISSA)	15	0.3	4,305	2.22	**-0.385**	**-0.084**	**-0.122**
sum(all violations)	25	0.5	5,727	2.96	**-0.356**	-0.022	**-0.101**

- Counts of annotations of coherence violations
- Correlation between violations and evaluation scores

Further Analysis

- Train linear regression to model relationship of particular errors to readability
- Most significant factors:
  - Missing/Misleading references
  - fragments
  - redundant content
  - poor coherence
- Total # of errors well-correlated with system ranks

Feature	Weight	Feature	Weight
Intercept	**3.407**	DNP-REF	**-0.157**
ACR-EXPL	-0.361	OTHRUNGR	**-0.155**
PRN+MISLA	**-0.355**	INCLDATE	**-0.151**
INCOMPLSN	**-0.275**	INP+REF	-0.067
NOSEMREL	**-0.262**	NODISREL	-0.046
REDUNDINF	**-0.259**	FM-EXPL	-0.023
PRN+MISSA	**-0.236**	SM+EXPL	0.038

Linear model weights

Friedrich et al, 2014, p. 1596, Tab. 3

Automatic Evaluation of Linguistic Quality

- Motivation
  - No focus on linguistic quality because no way to tune to it
  - Everyone uses ROUGE because you can tune
    - Explicitly tuned in many ML models
- Alternative strategies:
  - Micro — Learn to predict component scores
  - Macro — Learn to predict overall readability score
    - Intuitively: error count (LQVSumm) predicts well… but Errors manually derived

LQVSumm on Patas

- /dropbox/17-18/573/other_resources/LQVSumm

“Micro” Quality Prediction

Micro-Quality Prediction

- Pitler et al, 2010 via SVM Ranking
- Big Idea:
  - Train SVM classifier to compare whether one system is better than another
  - Use SVM to rank different system outputs

Language Quality Prediction Features
Discursive

- **Continuity**:
  - For each cohesive device, are sentences adjacent in source?
  - Position and confidence of antecedents of pronouns
  - Max, min, and average cosine similarity between sentences
- **Sentence fluency**
  - Shallow syntax features correlated w/MT quality
- Coh-Metrix (Online tool)
  - Set of psycholoinguistically-based coherence features + LSA similarity

Language Quality Prediction Features
Discursive

- **Word coherence**
  - cross-sentence word coocurrence patterns
- **Entity coherence**
  - via Entity-grids (Brown toolkit)

Language Quality Prediction Features
Syntactic

- **General word choice, sequence**
  - Language Models
- **Named Entities**
  - Modifiers for 1st mention of PERSON
  - Proportion of summary NER first mentions originall non-first
- **NP syntax**
  - POS, phrase tags in NPs

Language Quality Prediction Features
Syntactic

- **Local coherence devices counts:**
  - demonstratives
  - pronouns
  - definite descriptions
  - sentence initial discourse connectives

Results

System-level prediction accuracies

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	87.6	83.0	91.2	85.2	86.3
Named ent.	78.5	83.6	82.1	74.0	69.6
NP syntax	85.0	83.8	87.0	76.6	79.2
Coh. devices	82.1	79.5	82.7	82.3	83.7
Continuity	88.8	**88.5**	**92.9**	**89.2**	**91.4**
Sent. fluency	**91.7**	78.9	87.6	82.3	84.9
Coh-Metrix	87.2	86.0	88.6	83.9	86.3
Word coh.	81.7	76.0	87.8	81.7	79.0
Entity coh.	90.2	88.1	89.6	85.0	87.1
Meta ranker	**92.9**	87.9	91.9	87.8	90.0

Input-level prediction accuracies

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	66.3	57.6	62.2	60.5	62.5
Named ent.	52.9	54.4	60.0	54.1	52.5
NP Syntax	59.0	50.8	59.1	54.5	55.1
Coh. devices	56.8	54.4	55.2	52.7	53.6
Continuity	61.7	62.5	**69.7**	**65.4**	**70.4**
Sent. fluency	**69.4**	52.5	64.4	61.9	62.6
Coh-Metrix	65.5	**67.6**	67.9	63.0	62.4
Word coh.	54.7	55.5	53.3	53.2	53.7
Entity coh.	61.3	62.0	64.3	64.2	63.6
Meta ranker	**71.0**	**68.6**	**73.1**	**67.4**	**70.7**

Pitler et al, 2010 p. 550–551;Tab 2–3

Findings

- Overall accuracies quite good
  - ~70% on ranking, 90% pairwise
- Systems overall easier to rank than particular input
- **Continuity** related features best across components
  - Ensemble of ordering, coreference, cosine similarity cues
- Specifically tuned fluency scorer works on fluency

“Macro” Quality Prediction

Macro-Quality Prediction

Lin et al, (2012)

- High-level concept:
  - Discourse version of entity grid
    - Columns; entities (same head)
    - Rows: sentences
    - Cell values: PDTB Discourse Relation.Arg# tuples
  - Use linear programming to find optimal fit to discourse relations between sentences

Macro-Quality Prediction

Lin et al, (2012)

- Variants:
  - Inter-cell sequence frequencies
    - + Additional tuples:
      - Add "Explitic" / "Non-Explicit" tags to relation tuple
    - + Intra-cell bigrams
      - (Like "Role n-grams" from SOX paradigm)

S_1: Japan normally depends heavily on the Highland Valley and Cananea mines as well as the Bougainville mine in Papua New Guinea

S_2: Recently, Japan has been buying copper elsewhere.

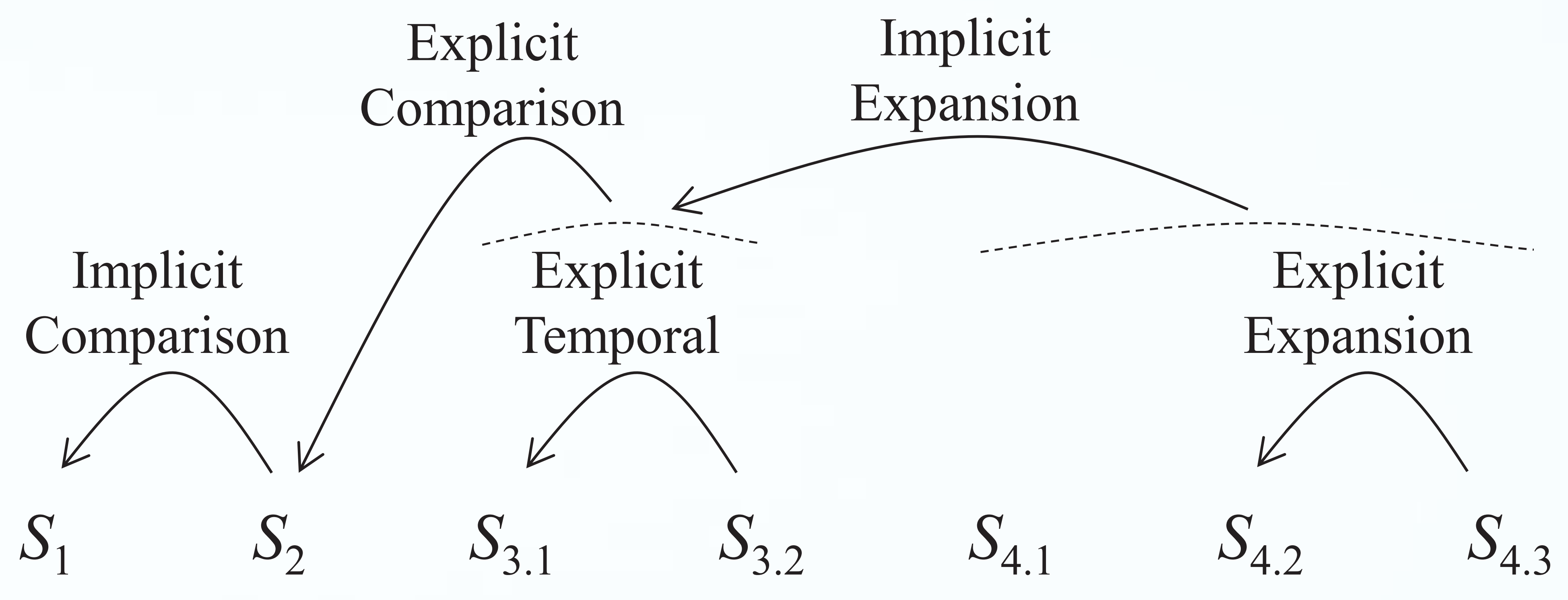
$S_{3.1}$: But as Highland Valley and Cananea begin operating,

$S_{3.2}$: they are expected to resume their roles as Japan's suppliers.

$S_{4.1}$: According to Fred Demier, metals economist for DBL, New York,

$S_{4.2}$: "Highland Valley has already started operating

$S_{4.3}$: and Cananea is expected to do so soon.



S #	Terms			
	copper	cananea	operat	depend
S_1	—	Comp.Arg1	—	Comp.Arg1
S_2	Comp.Arg2 Comp.Arg1	—	—	—
S_3	—	Comp.Arg2 Temp.Arg1 Exp.Arg1	Comp.Arg2 Temp.Arg1 Exp.Arg1	—
S_4	—	Exp.Arg2	Exp.Arg1 Exp.Arg2	—

Lin et al, (2012) p. 1010; Fig 1,2; Tab 2

Results

- Very strong correlations with manual readability score
- Beats prior predictors

Measure	Pearson	Spearman
ROUGE-2	0.7524	0.3975
TAC System 6	0.8194	0.4837
DiscRelGrid	0.8556	0.6593
DiscRelGrid +Explicit Tags + Within Cell Transcriptions	0.8666	0.7122