Content Realization #2

LING 573 — Systems & Applications

May 10, 2018

UNIVERSITY OF
WASHINGTON

PROFESSIONAL MASTER'S IN
COMPUTATIONAL LINGUISTICS

Begin Recording

UNIVERSITY OF
WASHINGTON

PROFESSIONAL MASTER'S IN
COMPUTATIONAL LINGUISTICS

Announcements:

- Please include your condor file as **../src/D3.cmd**

  - Helps with ease of grading having consistent location.

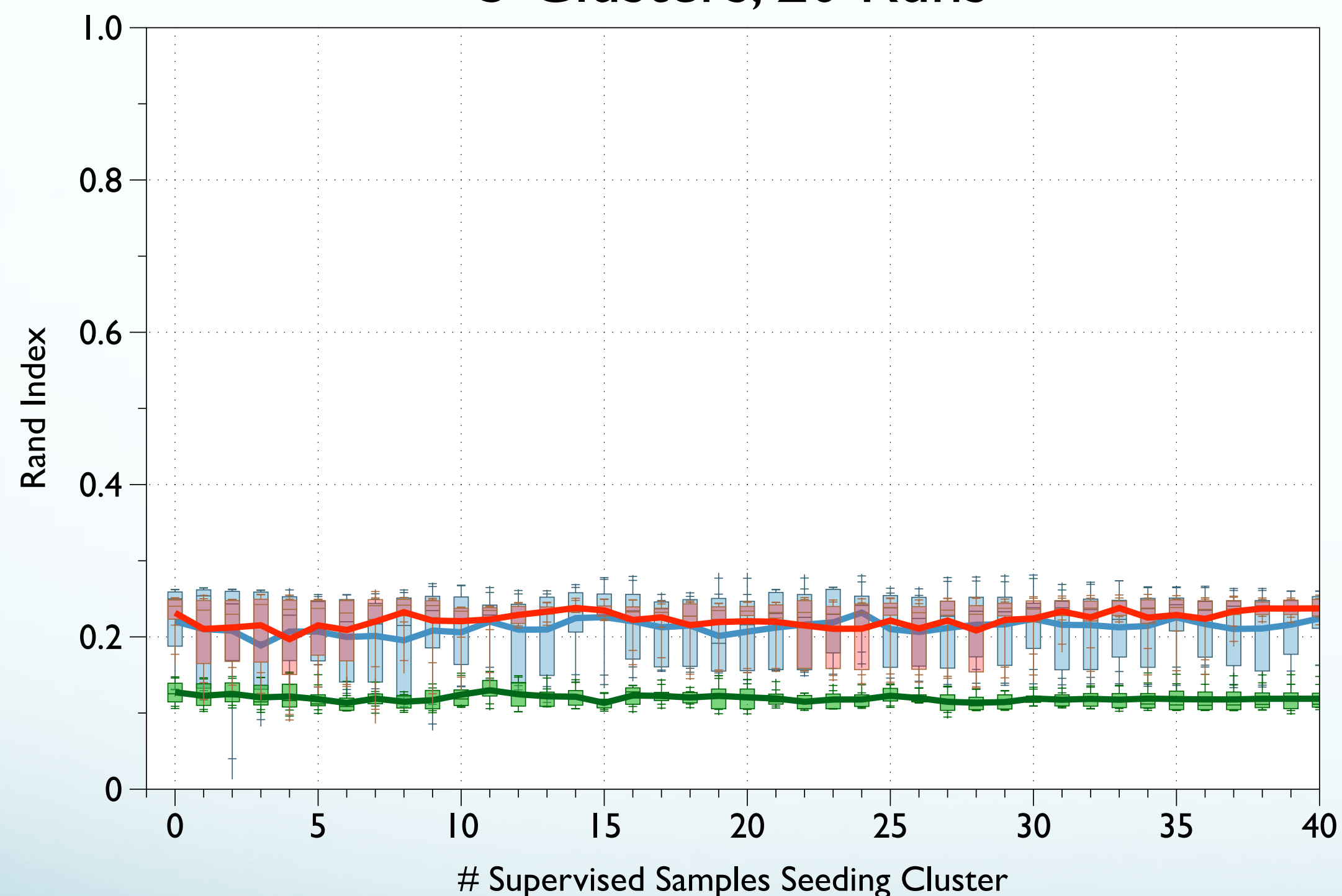# Miscellanea

I Feel Personally Attacked By XKCD

	COMMENT	DATE
	CREATED MAIN LOOP & TIMING CONTROL	14 HOURS AGO
	ENABLED CONFIG FILE PARSING	9 HOURS AGO
	MISC BUGFIXES	5 HOURS AGO
	CODE ADDITIONS/EDITS	4 HOURS AGO
	MORE CODE	4 HOURS AGO
	HERE HAVE CODE	4 HOURS AGO
	AAAAAAAA	3 HOURS AGO
	ADKFJSLKDFJSDKLFJ	3 HOURS AGO
	MY HANDS ARE TYPING WORDS	2 HOURS AGO
	HAAAAAAAAANDS	2 HOURS AGO

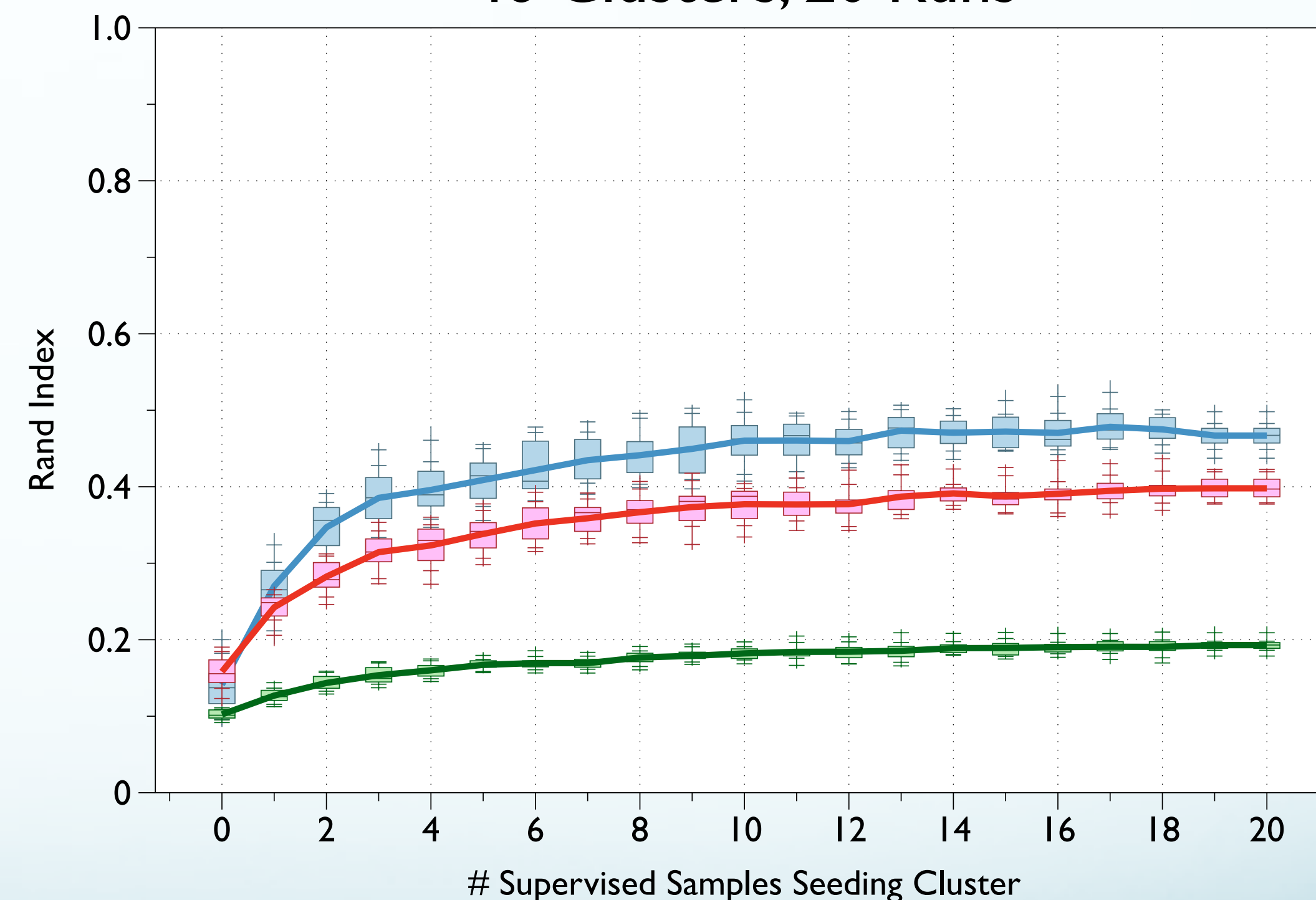AS A PROJECT DRAGS ON, MY GIT COMMIT MESSAGES
GET LESS AND LESS INFORMATIVE

Merge branch 'asdfasjkfdlas/alkdjf' into sdkjfls-final

The Continuing Saga of Embeddings

KMeans Clustering on TAC Categories
5 Clusters, 20 Runs

KMeans Clustering on TAC Topics
46 Clusters, 20 Runs

= 100M ENG_GW

= 10M ENG_GW

= 1M ENG_GW

6

The Continuing Saga of Embeddings

- gensim Doc2Vec embeddings on 1M, 10M, 100M:

  - /dropbox/17-18/573/other_resources/word_embeddings/eng_gw

- 1B, full set coming soon.

- Obviously, YMMV

Roadmap / Overview

- **So far**:

  - We've discussed Linguistic Quality

  - How to define the particular issues that affect summarization system

  - Examined a system meant to (re)rank output based on linguistic quality

- **Now**:

  - Improving references

    - Contributes to better readability,

  - Introduction to compression

    - Can help with long, run-on sentences — improve included content

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Content Realization:
Reference Resolution

NP-Rewrite
([Nenkova, 2008](#))

- Intuition:

  - Referring expressions are common sources of errors

  - References to people prevalent in news data, summaries

  - Targeted rewriting can improve readability

- Other intuitions:

  - Some entities are more important than others

  - Important entities may need less reintroduction

Challenges

- Lack of training data:

  - No summary data labeled for information status

- Readers sensitive to referring expressions

  - Prior work on NP rewriting has shown mixed results

    - Some improvement, some failures

- Relies on coreference

  - May contain errors

NP Rewrite
Very Good Example

- While the British government defended **the arrest**, it took no stand on extradition of Pinochet to Spain, leaving it to the courts.

- While the British government defended **the arrest in London of former Chilean dictator Augusto Pinochet**, it took no stand on extradition of Pinochet to Spain, leaving it to British courts.

NP Rewrite
Mixed Example

- **Duisenberg** has said growth in the euro area countries next year will be about 2.5 percent, lower than the **3 percent** predicted earlier.

- **Wim Duisenberg**, **the head of the new European Central Bank**, has said growth in the euro area countries next year will be about 2.5 percent, lower than **just 1 percent in the euro-zone unemployment** predicted earlier.

NP-Rewrite: Approach
(Nenkova, 2008)

1. Estimate "importance" of each content word by MLE

$$p(w_i) = \frac{n_i}{N}$$

2. Estimate weight for sentence by sum of $p(w)$s

3. Select **sentence with highest weight**

   a. Find all **NPs in this sentence**

   b. Find **highest-weighted coreferring NPs** across all documents

   c. Replace **NPs in this sentence** with **highest-weighted NP**

Results

- Not significant improvement

  - 0.4039 — 0.4169 on Pyramid

  - Out of 50 test cases:

    - 22 worse

    - 23 better

    - 5 the same

Discussion

- **Problems**:

  - Extremely basic content selection approach

  - Relies on coreference — difficult task to perform automatically

- **Possible Improvements**:

  - Perhaps similar method could be used with more advanced content selection

Information Status Distinctions:
([Siddarthan et al 2011](#))

- **Intuition**:

  - Discourse and background world knowledge are proxies for "**Information Status**"

  - Referring expressions should be conditioned upon their information status

  - Information status constrains realization

- **Approach**:

  - Automatically identify information status

  - Use **IS** to guide rule-based generation of referring expressions

Information Status Distinctions:
(Siddarthan et al 2011)

- **discourse-new** vs. **discourse-old**

  - Is this **first mention** of person, or **subsequent mention**?

- **hearer-new** vs. **hearer-old**

  - Person **less well known** vs. likely **very well known**

  - Don't waste space describing well-known individuals

- **Major** vs. **Minor** Character

  - Salience of the person in the news story

Information Status Distinctions:
(Siddarthan et al 2011)

- Research Question:
    - Is it possible to infer these distinctions via modeling?
    - Apply this model to predict information status?

Corpus Analysis

- Assess relation between:

  - Information status and referring expressions

- Used data from DUC 2001–2003

- NER, Coreference automatic

		Discourse-new	Discourse-old
Name Form	Full name	0.97	0.08
	Surname only	0.02	0.87
	Other (e.g., Britney, JLo)	0.01	0.05
Pre-Modification	Any	0.51	0.21
	None	0.49	0.79
Post-Modification	None	0.60	0.89
	Apposition	0.25	0.04
	Relative clause	0.07	0.03
	Other	0.08	0.04
Any Modification (Either Pre- or Post-)	Some Modification	0.76	0.30
	No Modification	0.24	0.70

Siddarthan et al, 2011 p. 818, Table 1

Generating Discourse-New

1. **IF** NP head is person name,

    a. **IF** appears with pre-modifier, **THEN**

      i. insert longest pre-modifier + full name

    b. **ELSE IF** appears with apposition modifier **THEN**

      i. add apposition to reference

2. **ELSE** don't rewrite

Generating Discourse-Old

1. Use surname only, remove all pre/post modifiers.

Example Rewrite

- **Honecker** has come under investigation for charges of corruption and living in luxury at the cost of the state. **Former East German leader Erich Honecker** may be moved to a monastery to protect him from a possible lynching by enraged citizens. As protests gathered strength last fall, **Erich Honecker**, **East Germany's longtime orthodox leader** "lost touch with reality," according to the man who succeeded him as Communist leader only to be ousted later. **Ousted East German leader Erich Honecker**, who is expected to be indicted for high treason, was arrested Monday morning upon release from a hospital and taken to prison.

- **Former East German leader Erich Honecker** has come under investigation for charges of corruption and living in luxury at the cost of the state. **Honecker** may be moved to a monastery to protect him from a possible lynching by enraged citizens. As protests gathered strength last fall, **Honecker** "lost touch with reality," according to the man who succeeded him as Communist leader only to be ousted later. **Honecker**, who is expected to be indicted for high treason, was arrested Monday morning upon release from a hospital and taken to prison.

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Detecting Features

- **discourse-new** vs. **discourse-old**:

    - Obvious from summary order

- **Major** vs. **Minor** and **hearer-new** vs. **hearer-old?**

    - How to establish hearer or major/minor status?

    - Train on mentions in model summaries

Detecting Features

- **hearer-new** vs. **hearer-old**

  - Infer familiarity via whether model summaries include:

  - **title/role+surname** or **unmodified fullname**

- **Major** vs. **Minor**:

  - **Referred to by name** in some human summary of topic vs. **never referenced**

  - 258 major/3926 minor by data

Classifier Training: Features

- **frequency**:
  - "rf•idf" for references (number of references, total docs w/reference)
  - Proportion of first mentions containing full name
  - Number times person referenced after first mention

- **lexical**
  - Probability of first reference by bigram
  - Top 20 high frequency description words present in initial references (est. from newswire)

Classifier Training: Features

- **syntactic**

  - # appositives or rel. clauses attaching to initial reference

  - # times appositives used

  - # times rel. clauses used

  - # copula constructions involving referent

- **classifiers:**

  - Used SVM, Decision Trees

Classifier Results

System	All Features	Majority Class Baseline
hearer-new	0.75	0.70
hearer-old	0.76	0.00
major	0.60	0.00
minor	0.98	0.97

- Significantly better than baseline in all cases

Applying Classifier Decision

IF discourse-new AND NP head is person name

    IF minor

        exclude name, use only role, modifiers, etc

    IF MAJOR and **hearer-old**

        Include name and role/temporal (only)

    IF MAJOR and **hearer-new**

        Include name and role/temporal

        Also include affiliation, post-mod (classifier)

IF discourse-old:

    Surname ONLY

Evaluation

- Created (nearly) deterministic rule set

  - Based on information status classification

  - To rewrite referring expressions in extractive summaries

- Evaluated in **paired preference tests**

  - Compared **original extractive** and **rewritten** summaries

- Where a preference was expressed

  - Rewritten summaries rated as more coherent

  - **Extractive** rated as *more informative*

    - Why? Rewrite rules generally shrink rather than add content

Discussion

- Pros:

  - Intuitive, interpretable model

  - Solid results: ~0.75 accuracy, higher if humans agree

  - Often preferred to extract

- Cons:

  - Limited: only applies to person names

  - Error propagation: coreference, NP extraction

  - Ignores other aspects of realization, i.e. length

UNIVERSITY OF
WASHINGTON

PROFESSIONAL MASTER'S IN
COMPUTATIONAL LINGUISTICS

Sentence Compression

UNIVERSITY OF
WASHINGTON

PROFESSIONAL MASTER'S IN
COMPUTATIONAL LINGUISTICS

Sentences Compression Motivation

- Summaries are short!

- Maximizing information means removing extraneous information.

- Former East German leader Erich Honecker, who "lost touch with reality,"…

  - → Honecker…

    - …saving of ten words!

    - Remove anaphora

    - Remove relative clause

Sentences Compression Motivation

- *As U.S. Court of Appeals Judge Sonia Sotomayor made her Senate debut with a series of private meetings, Republicans said they would prefer holding hearings on her nomination in September, which could cloud the speedy summertime confirmation Obama wants.*

- In this sentence, there are:

  - Discussion of the series of meetings

  - Timing of meetings

  - Political opinion of meetings

Compression Approaches

- At a high level:

  - Find ways to remove portions of sentences

  - …while maintaining grammaticality!

- Approaches:

  - Heuristc / Rule-Based

  - Machine Learning Approaches

  - Integrated with Selection / Optimization

Overview of Approaches

- Heuristic approaches

  - Deep vs. Shallow processing

  - Information- vs. readability-oriented

- Machine-learning approaches

  - Sequence models: HMM, CRF

  - Deep vs. Shallow information

- Integration with selection

  - Pre/post-processing: Candidate selection: heuristic/learned

Common to Many Approaches

- Start with syntactic parse

  - Way to represent sentence as subcomponents

  - Can learn or write rules

    - For which parse rules / subtrees should be maintained or not

Form	CLASSY	ISCI	UMd	SumBasic+	Cornell
Initial Adverbials	Y	M	Y	Y	Y
Initial Conj	Y		Y	Y	
Gerund Phrases	Y	M	M	Y	M
Relative clause appositives	Y		M	Y	Y
Other adv	Y				
Numeric: ages,	Y				
Junk (byline, edit)	Y				Y
Attributives	Y	Y		Y	Y
Manner modifiers	M	Y	M		Y
Temporal modifiers	M	Y	Y		Y
POS: det, that, MD			Y		
XP over XP			Y		
PPs (w/, w/o constraint)			Y		
Preposed Adjuncts			Y		
SBARs			Y		M
Conjuncts			Y		
Content in parentheses		Y			Y

After Presentations:

- Deeper dive into sentence compression approaches

Deliverable #3

- **Code** — Due Monday, 11:00pm PT

- **Report** — Due Thursday, 11:00pm PT

- **Presentations** — Due 1pm day of presentations

  - Presentation Time Slots:

    - Will be randomized—will post this afternoon

    - Will reverse order for D4.

- **Note!**

  - Please include your condor submit file in **../src/D3.cmd**.

Deliverable #4

- Final system

  - Continue system improvement

  - Add content realization

- **Evaluation**:

  - Devtest (2010)

  - Evaltest (2011)

    - New blind test