# Optimization, Neural Approaches, Information Ordering

LING 573 — Systems & Applications

Ryan Georgi — 4/17/2018

*Begin Recording!*

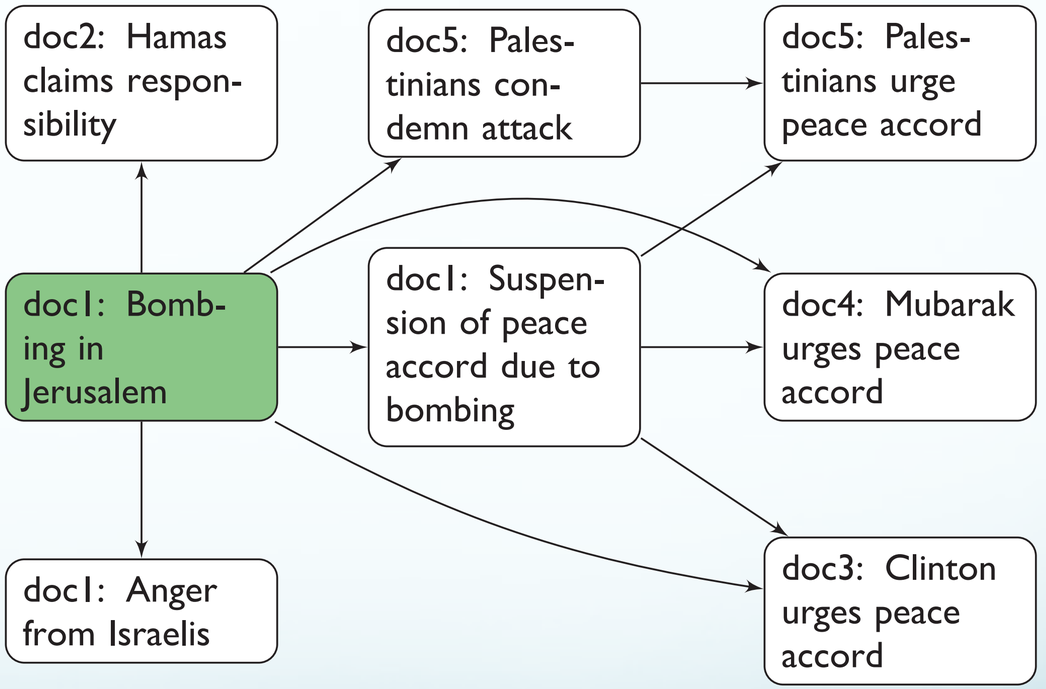
# One More Discourse Approach!

- Christensen et. al, 2013, 2014 — GFLOW
  - Rather than perform deep discourse analysis, find simple **ordering constraints**
    - Look for "*but*," "*however*," "*moreover*"
    - Also look for coreference
  - Build multi-document discourse ***directed*** graph of sentences
    - Use number of indicators as weights

# GFLOW

## (Christensen et. al, 2013, 2014)

- Example discourse graph
- Intuition:
  - Coherent summary begins with inciting event
  - Reactions follow



# GFLOW

## (Christensen et. al, 2013, 2014)

- **Preprocessing:** Coreference, Deverbal Noun Reference ('bombing' vs. 'attacked')
- **Salience —** Statistical approaches (LLR, tf*idf)
- **Redundancy** — handled by comparing Verb(arg1,arg2) tuples
- **Coherence** — as calculated by weighted graph

**maximize:**

$$F(x) \triangleq Salience(X) + \alpha \cdot Coherence(X) - \beta \mid X \mid$$

- $\beta|X|$ is brevity penalty

# GFLOW: Results

## (Christensen et. al, 2013, 2014)

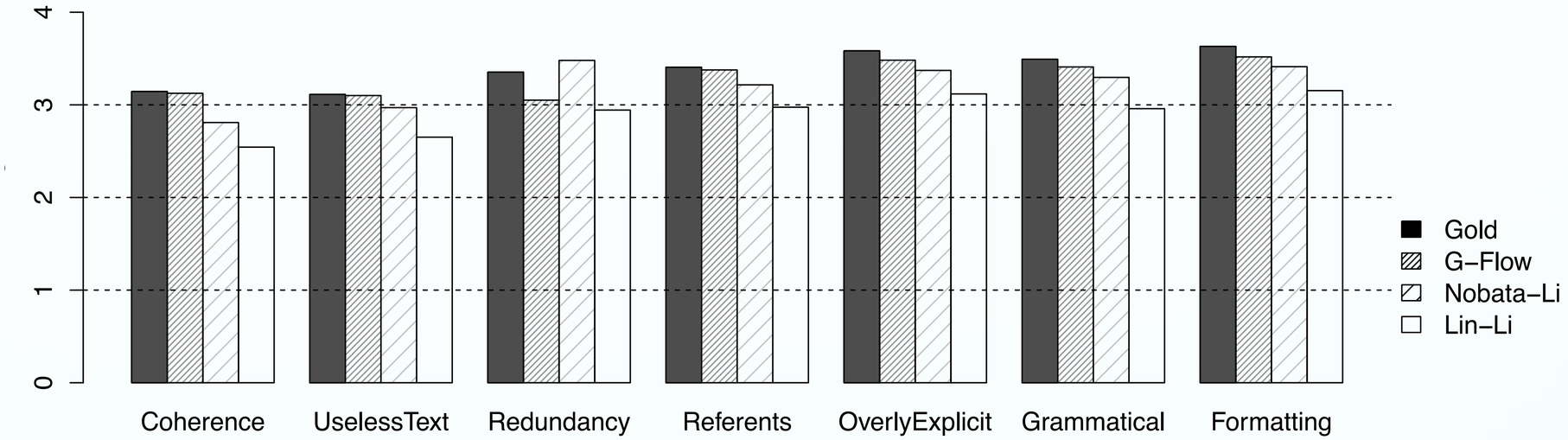
- Does not beat state-of-the-art system in ROUGE metrics
- But what about others?
- Note: Comparing one gold standard summary against others is also not particularly high ROUGE-1

| System | R | F |
|---|---|---|
| NOBATA | 30.44 | 34.36 |
| Best system in DUC-04 | 38.28 | 37.94 |
| Takamura and Okumura (2009) | 38.50 | - |
| LIN | 39.35 | 38.90 |
| G-FLOW | 37.33 | 37.43 |
| Gold Standard Summaries | **40.03** | **40.03** |

ROUGE-1 Recall and $F_1$ on DUC '04 Data

# GFLOW: Results

## (Christensen et. al, 2013, 2014)




- G-Flow better in all but Redundancy on manual evaluations (via AMT workers)

# GFLOW: Takeaways

## (Christensen et. al, 2013, 2014)

- Discourse can be helpful
  - but perhaps more in information ordering
  - Promotes coherence

# Summarization as Optimization

# Key Concept

- Extractive summarization can be thought of as *global inference problem*
- Best summary is a "solution" in the search space that:
  - Maximizes **relevance**
  - Minimizes **redundancy**
  - Is bounded in **length**

# Global Inference

- Many approaches we've looked at optimize separately
  - One process to **maximize relevancy**
  - One process to **minimize redundancy**
- Many approaches to global optimization problem
  - One is Integer Linear Programming (ILP)

# Integer Linear Programming (ILP)

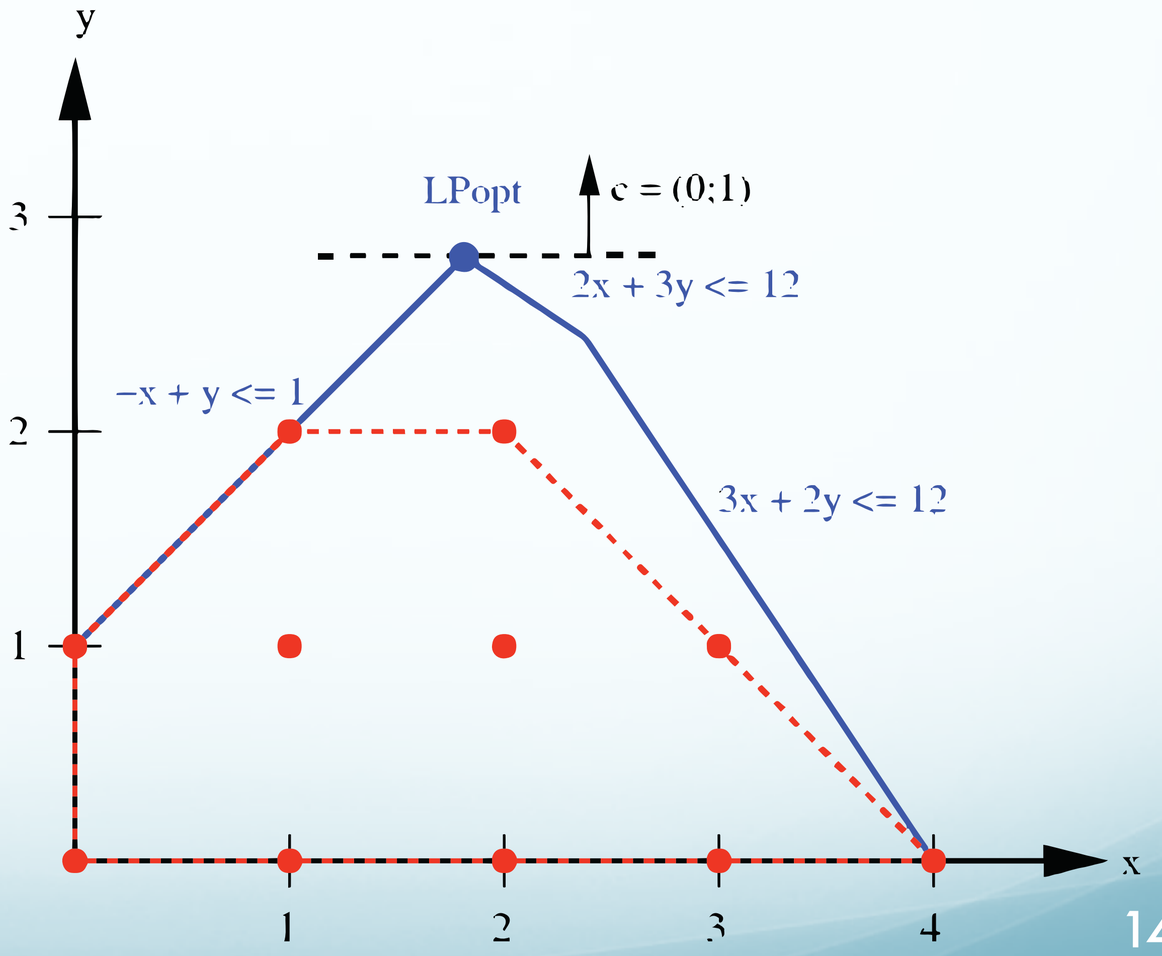
# Integer Linear Programming

- A constrained subtype of optimization problem
- An integer linear program specifies
  - A single linear maximization term
  - Subject to linear equality/inequality constraints
  - Involving integer valued variables
- Free ILP Toolkits Available
  - GNU Linear Programming Kit (GLPK)
  - Examples on Github with scipy.optimize

# Integer Linear Programming

- Example: find $\max y$

$$-x + y \leq 1$$
$$3x + 2y \leq 12$$
$$2x + 3y \leq 12$$
$$x, y \geq 0$$
$$x, y \in \mathbb{Z}$$




- Solution is to find highest point **y** that obeys all constraints
  - Finds convex point that matches line representing objective function

# Summarization as ILP

- **For summarization:**
  - Map summary requirements to ILP elements
- **Summary goal**:
  - "Best" summary
- **Summary requirements**:
  - Minimize redundancy
  - Within desired length

# Summarization as ILP

(via McDonald, 2007)

- maximize: $\sum_i \alpha_i Relevancy(i) - \sum_{i<j} \alpha_{ij} Redundancy(i,j)$
- Such that $\forall i,j$:

(1) Sent $\alpha_i$ or sent pair $\alpha_{ij}$ are included or not — $(1)\ \alpha_i, \alpha_{ij} \in \{0,1\}$

(2) Length of all sentences must be ≤ total length $K$ — $(2)\ \sum_i \alpha_i l(i) \leq K$

(3) If sent pair $\alpha_{ij}$ is included, $\alpha_i$ must be included — $(3)\ \alpha_{ij} - \alpha_i \leq 0$

(4) If sent pair $\alpha_{ij}$ is included, $\alpha_j$ must be included — $(4)\ \alpha_{ij} - \alpha_j \leq 0$

(5) If $\alpha_i$ and $\alpha_j$ are included, sent pair $\alpha_{ij}$ must be included — $(5)\ \alpha_i + \alpha_j - \alpha_{ij} \leq 1$

# Optimization Approaches to Reducing Redundancy

- DPP: Determinantal Point Processes [python GH] (Kulesza & Taskar 2012)
  - Set models balancing information importance w/diversity
- ICSISumm: Uses Integer Linear Programming frame [code] (Gillick et al, 2008)
  - Optimizes coverage of key bigrams weighted by document frequency
- OCCAMS_V (Davis et al, 2012)
  - Uses LSA (Latent Semantic Analysis) to weight terms
  - Sentence selection via optimization problems:
    - Budgeted maximal coverage; knapsack

# ICSISumm

- Key ideas:
  - Cast summarization as optimization problem
  - Identify important "concepts" to incorporate
  - Build best such summary
  - Implemented as Integer Linear Programming

# Representing Concepts

- Concepts = Bigrams
  - Stemmed
  - No stopword-only bigrams
  - Occuring in at least 3 documents
- Weights
  - Document frequency
  - # Of Documents (from cluster) for bigram
- Selected sentences must contain ≥ 2 query terms

# Results

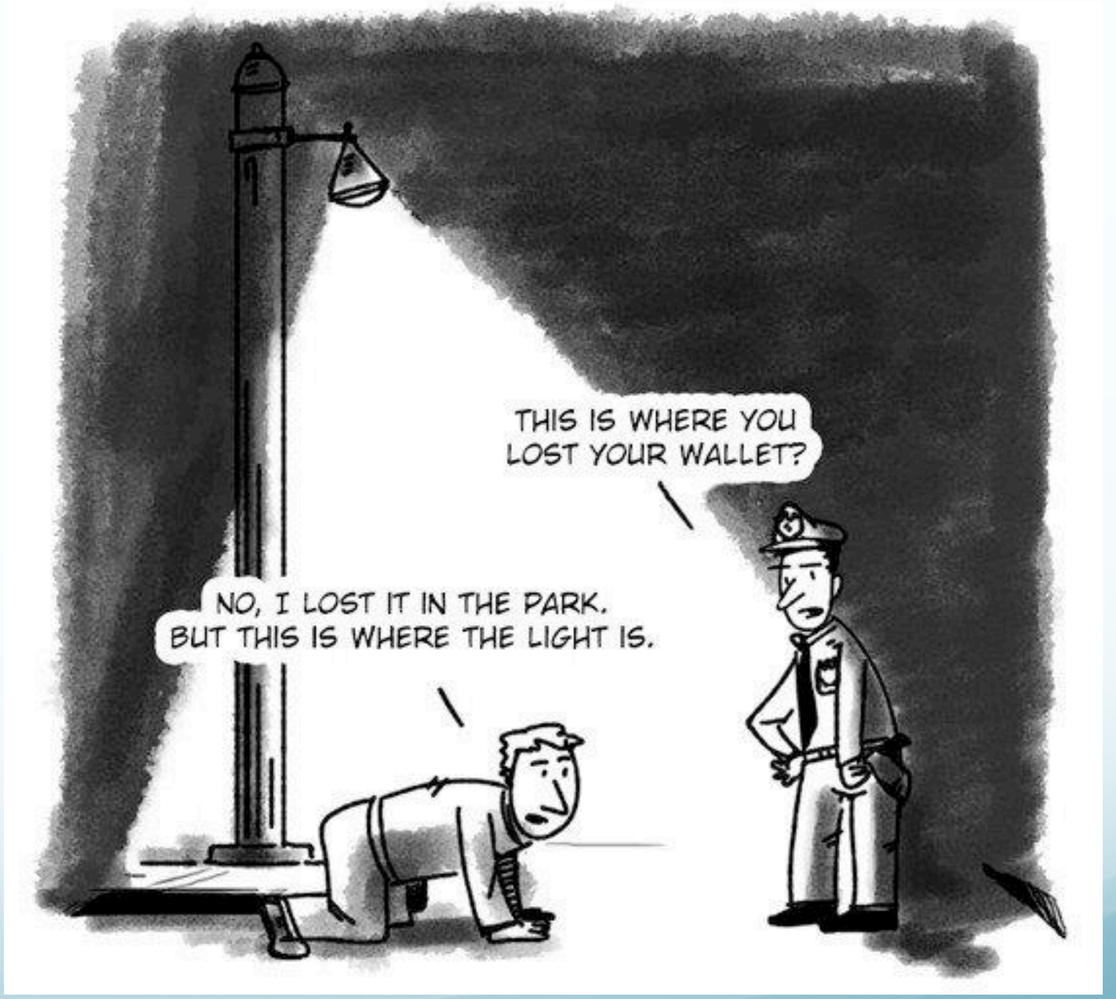
- ICSISumm, TAC 2008
- After using open source solver
- 2009 results:
  - 2nd best pyramid, ROUGE-2
  - Best ROUGE-3, ROUGE-4

| Metric | ICSI-1 (R) | ICSI-2 (R) | Best |
|---|---|---|---|
| Resp | 2.689 (9) | 2.238 (28) | 2.792 |
| Ling | 2.479 (21) | 2.021 (47) | 3.250 |
| Pyr | 0.345 (2) | 0.324 (10) | 0.362 |
| R-1 | 0.379 (5) | 0.383 (4) | 0.391 |
| R-2 | 0.110 (2) | 0.111 (1) | – |
| R-3 | 0.049 (1) | 0.044 (2) | – |
| R-4 | 0.023 (1) | 0.022 (3) | – |
| R-SU4 | 0.134 (4) | 0.143 (1) | – |
| BE | 0.063 (2) | 0.064 (1) | – |

# NN Approaches

# NN Approaches Overview

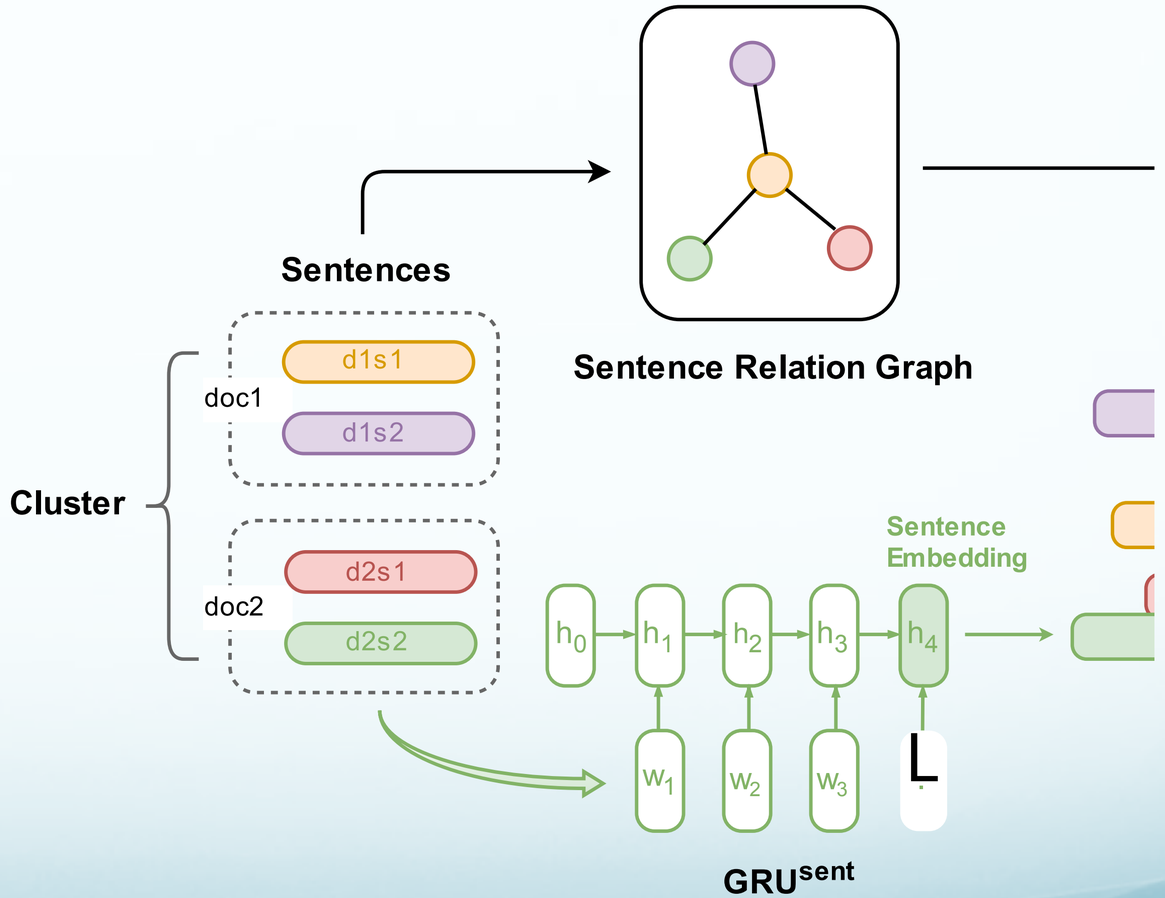
- Most work done so far is on single-document summarization
- Lots of interest in abstractive approaches
  - Large amount of abstractive data available
    - CNN/DailyMail corpus [gh] (dropbox: `./other_resources/cnn-dm`)
    - TL;DR reddit summaries [data] [paper] (dropbox: `./other_resources/tldr`)
  - Lots of abstractive approaches



# Graph-Convolutional NN Summarization

(Yasunaga et. al, 2017)

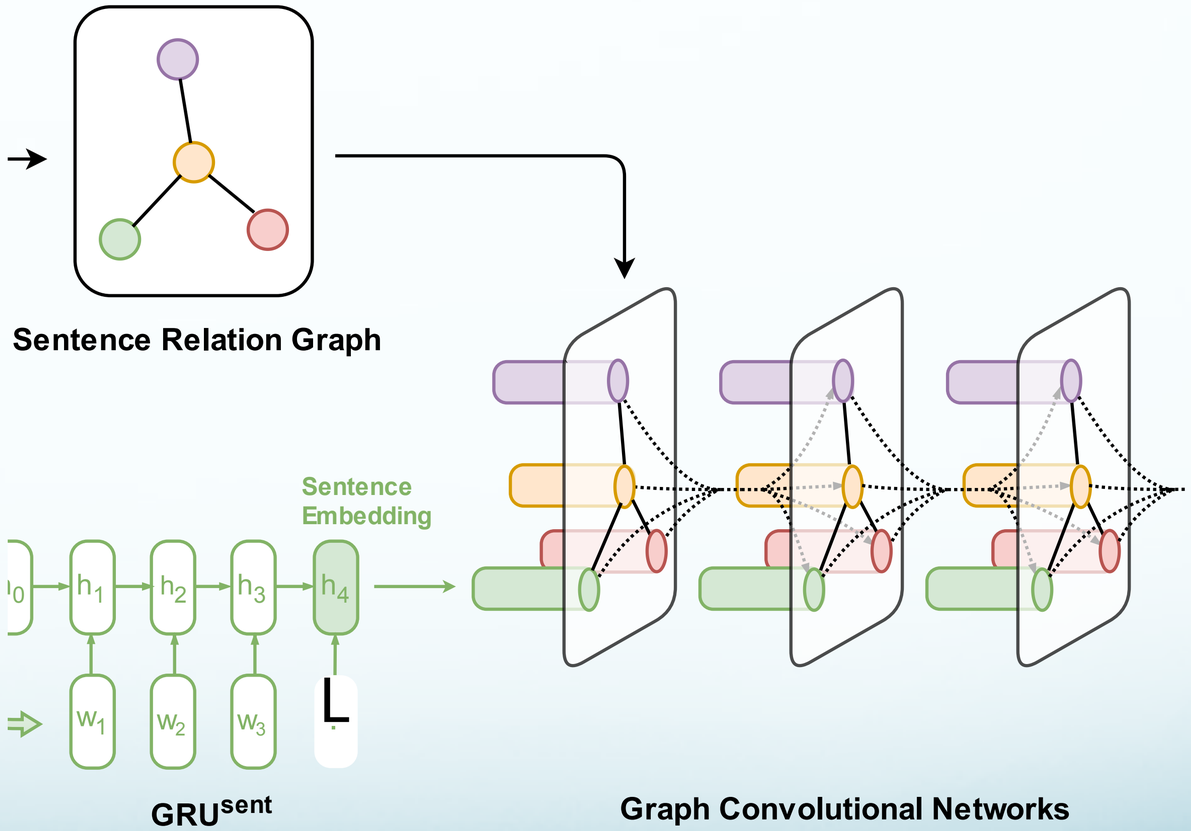
- Build sentence relation graph
  - Represent nodes in the graph with GRU-generated sentence embeddings



# Graph-Convolutional NN Summarization

(Yasunaga et. al, 2017)

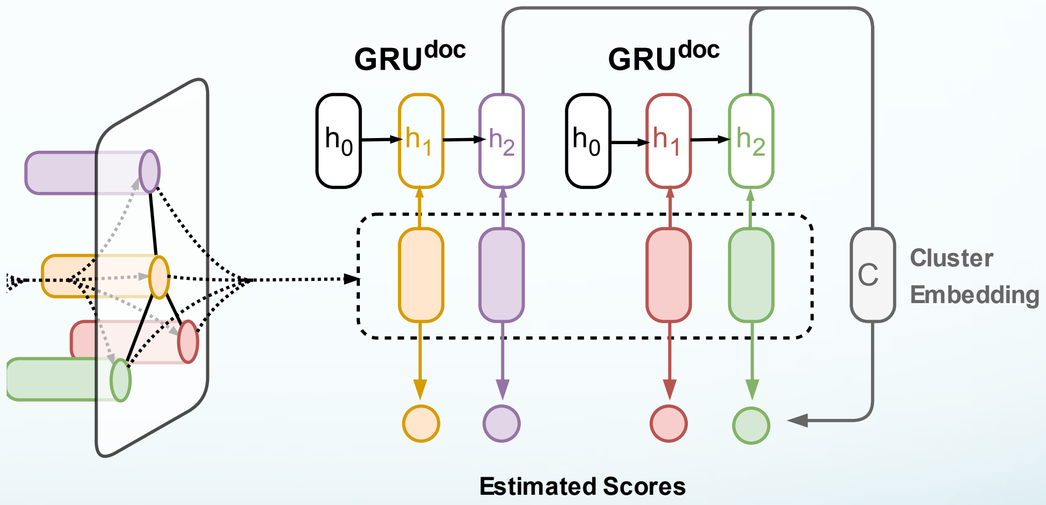
- Use Graph Convolutional Networks (Kipf & Welling, 2017)
  - Convolves (applies filters) to sent embeddings to produce 2nd order sent embeddings
  - Purpose of GCN — learn important features of graph (think: eigenvector)



# Graph-Convolutional NN Summarization

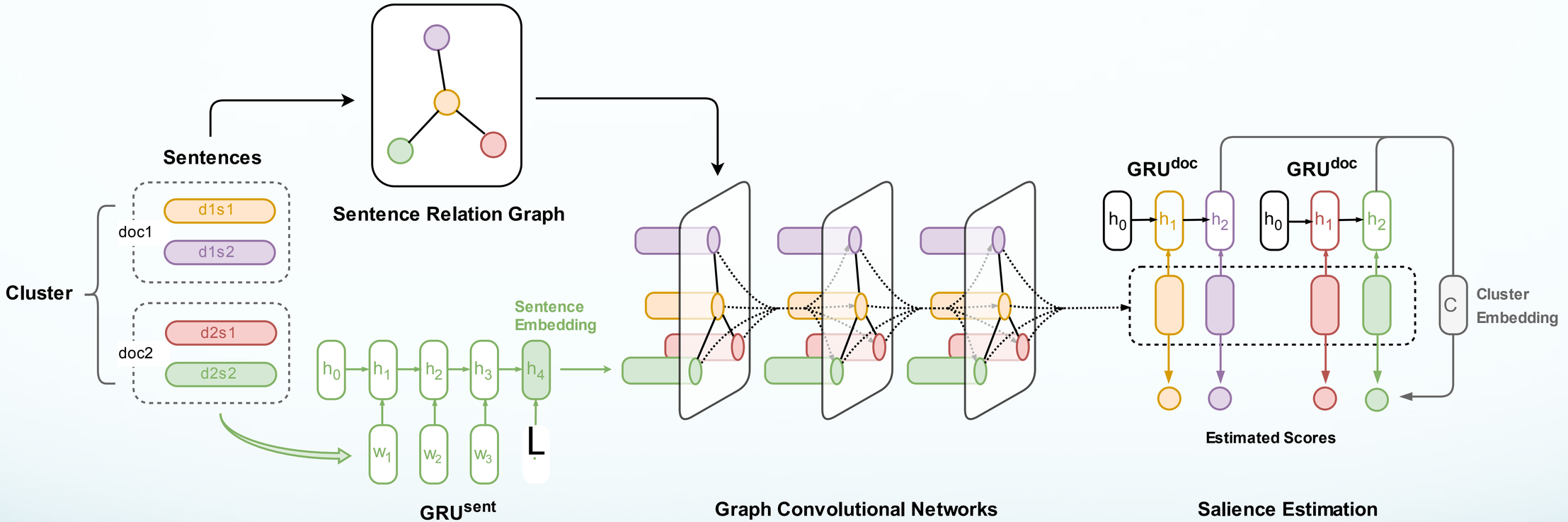
(Yasunaga et. al, 2017)

- Use GRU of sentence embeddings to represent clusters as cluster embeddings
  - (Cluster embedding ≈ neural-net-speak for centroid)
- Calculate salience by comparing sentence embedding to cluster embedding
- Pick salient sentences greedily



# Graph-Convolutional NN Summarization

(Yasunaga et. al, 2017)




(GRU = Gated Recurrent Unit)

# Graph-Convolutional NN Summarization

(Yasunaga et. al, 2017)

- Multiple graph-building approaches used
  - Cosine similarity
  - Approximate Discourse Graph (ADG)
    - (From GFLOW)
  - Personalized Discourse Graph (PDF)

# Results (DUC 2004 Eval Set)

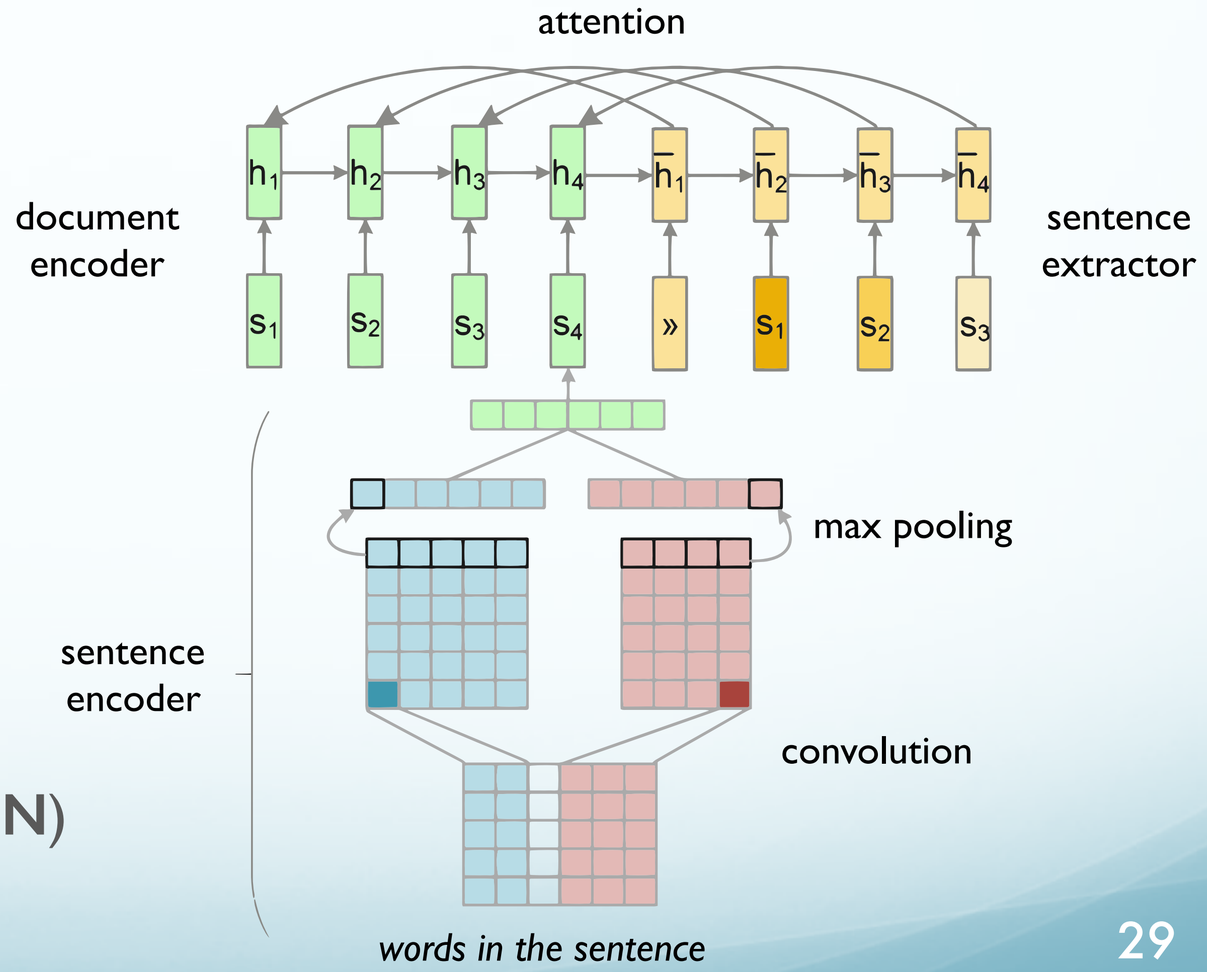
- "traditional" methods still competitive
- Does not use model summaries

| | R-1 | R-2 |
|---|---|---|
| SVR ( Li et al. , 2007) | 36.18 | 9.34 |
| CLASSY11 ( Conroy et al, 2011) | 37.22 | 9.20 |
| CLASSY04 ( Conroy et al, 2004) | 37.62 | 8.96 |
| GreedyKL (Haghighi and Vanderwende , 2009) | 37.98 | 8.53 |
| TsSum (Conroy et al, 2006) | 35.88 | 8.15 |
| G-Flow (Christensen et al, 2013) | 35.30 | 8.27 |
| FreqSum (Nenkova et al., 2006) | 35.30 | 8.11 |
| Centroid (Radev et al., 2004b) | 36.41 | 7.97 |
| Cont. LexRank (Erkan and Radev , 2004) | 35.95 | 7.47 |
| RegSum ( Hong and Nenkova, 2014) | **38.5** | **9.75** |
| GRU | $36.64_{\pm 0.11}$ | 8.47 |
| GRU+GCN: Cosine Similarity Graph | $37.33_{\pm 0.23}$ | 8.78 |
| GRU+GCN: ADG from G-Flow | $37.41_{\pm 0.32}$ | 8.97 |
| GRU+GCN: Personalized Discourse Graph | **38.23** $_{\pm 0.22}$ | **9.48** |

# Single Document, Extractive Approach

(Cheng & Lapata, 2016)

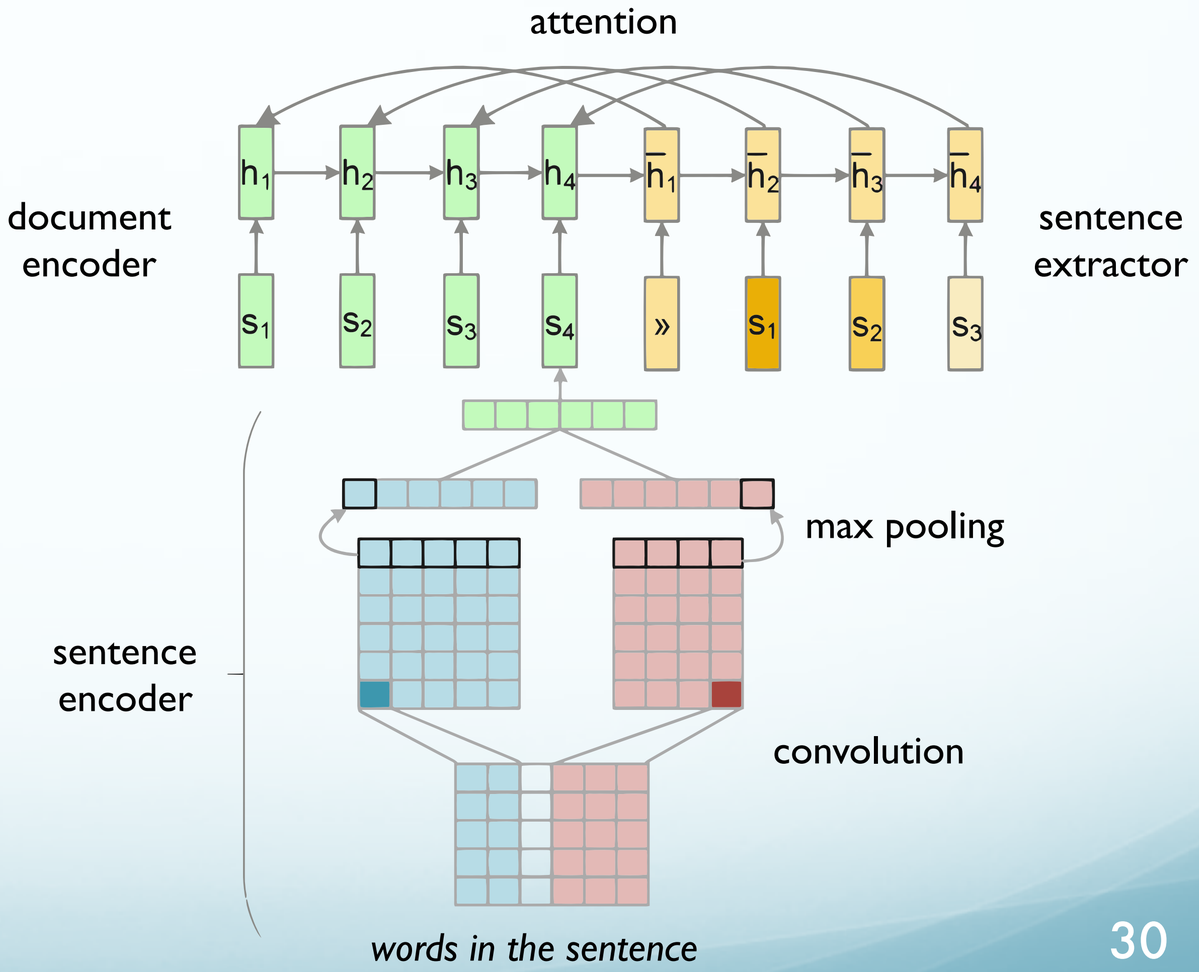
- Essentially similar to seq2seq
  - But **attention** used for extraction rather than focus for translation
- CNNs quite good at classification
- RNNs good at order information
- Combining both:
  - sentences classified by salience (**CNN**)
  - ordering of attention used for selection (**RNN**)



# Single Document, Extractive Approach

(Cheng & Lapata, 2016)

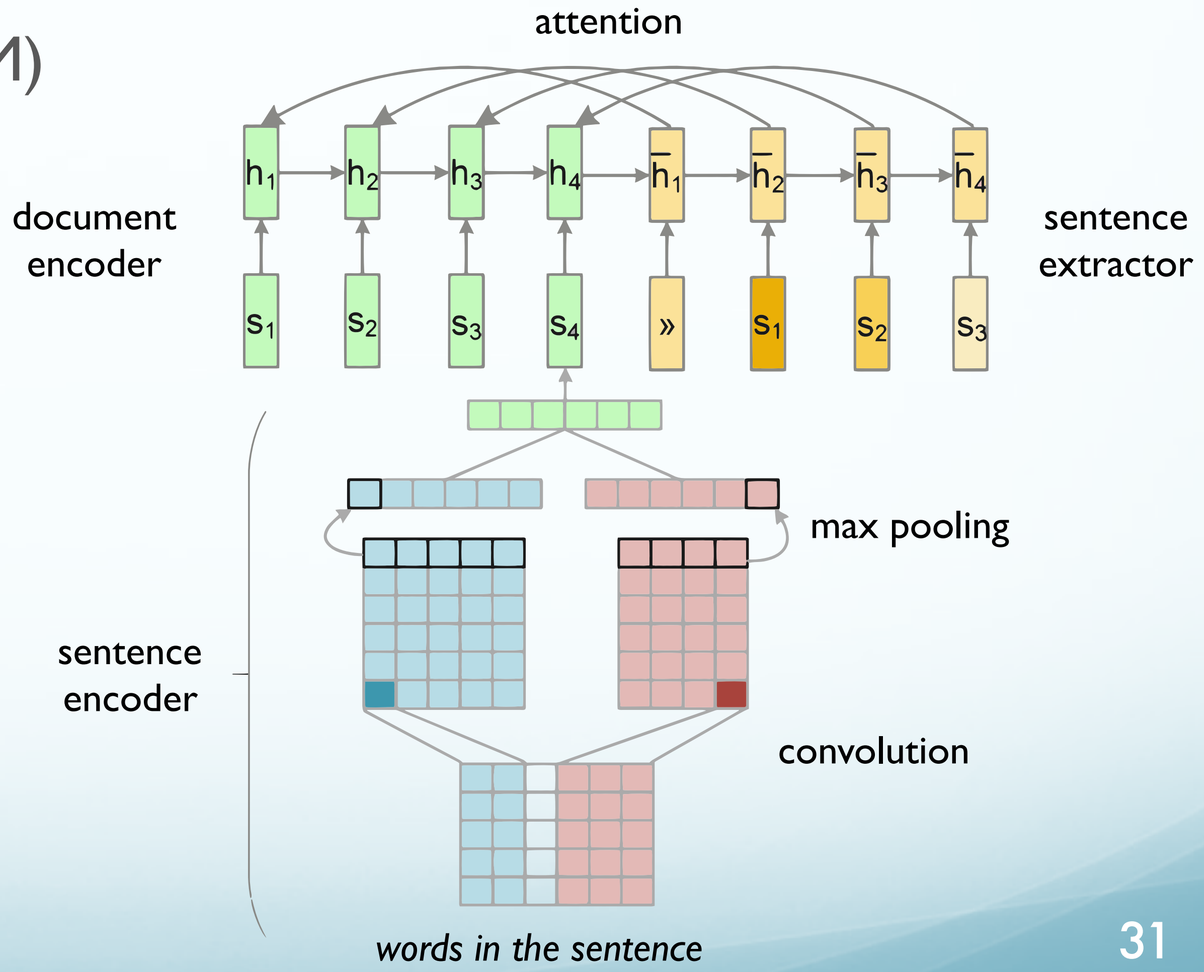
- **Training data**: DailyMail news highlights
  - Highlights = abstractive
  - Used highlights to determine doc sentences
- Methodology:
  - Used word embeddings for word input
  - Convolutions:
    - words→6-dim sentence embeddings
  - Max pooling to select salient convolved feats
  - Sentence embeddings input to RNN



# Single Document, Extractive Approach

(Cheng & Lapata, 2016)

- Sentence embeddings input to RNN (LSTM)
- Attention in LSTM used to highlight sentences for extraction
  - Conditioned upon previous decisions



# Single Document, Extractive Approach

(Cheng & Lapata, 2016)

- Outperforms LEAD and Logistic regression (LREG)
- Constraint-based approaches do better
  - ILP / uRank

| DUC 2002 | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| LEAD | 43.6 | 21.0 | 40.2 |
| LREG | 43.8 | 20.7 | 40.3 |
| ILP | 45.4 | 21.3 | 42.8 |
| NN-ABS | 15.8 | 5.2 | 13.8 |
| TGRAPH | 48.1 | 24.3 | — |
| URANK | 48.5 | 21.5 | — |
| NN-SE | 47.4 | 23.0 | 43.5 |
| NN-WE | 27.0 | 7.9 | 22.8 |

Eval on DUC-2002 Single-Document

# Information Ordering

# Information Ordering

- Basic Approaches:
  - Variations on chronological ordering
- Ensembles for ordering

# Basics

- Content Selection:
  - Identified sentences or information units for summary
- Information Ordering
  - Linearize selected content into smooth-flowing text

# Information Ordering: Factors

- Semantics
  - Chronology — respect the sequential flow of content (esp. events)
  - Discourse:
    - Cohesion — Adjacent sentences talk about the same thing
    - Coherence — Adjacent sentences naturally related

# Single vs. Multi-Document

- Single-document summarization: **Just keep original order.**
  - Chronology? *Ok*.
  - Cohesion? *Ok*.
  - Coherence? *Meh…*
- Multi-document
  - What does "original order" mean?
  - Chronology?
    - **Publication** order? Or **document-internal** order?
    - Differences in document ordering of information

# Information Ordering: A Bad Example*

*(The 2001 death of Ernest Hemingway's trans daughter, Gloria [formerly Gregory] Hemingway)*

1. Hemingway, 69, died of natural causes in a Miami jail after being arrested for indecent exposure.
2. A book [s]he wrote about [her] father, "Papa: A Personal Memoir," was published in 1976.
3. [She] was picked up last Wednesday after walking naked in Miami.
4. "He had a difficult life." [sic]
5. A transvestite who later had a sex-change operation, [she] suffered bouts of drinking, depression, and drifting, according to acquaintances.
6. "It's not easy to be the son of a great man," [sic] Scott Donaldson told Reuters.

*[*editorial brackets mine]*

# A Basic Approach

- Publication Chronology
- Given a set of ranked, extracted sentences…
- Order by:
  - Across articles
  - By publication date
- Then:
  - Within articles
- Clearly not ideal, but at least a reproducible baseline.

# Improving Ordering

- Improve some set of chronology, cohesion, coherence
- Chronology, cohesion (Barzilay et al, 2002)
- Key ideas:
  - Summarization and chronology over "themes"
  - Identifying cohesive blocks within articles
  - Combining constraints for cohesion within time structure

# Importance of Ordering

- Analyzed DUC summaries scoring poor on ordering
- Manually reordered existing sentences to improve
- Human judges scored both sets:
  - Incomprehensible, Somewhat Comprehensible, Comprehensible
- Manually reorderings ***as good or better*** than originals
- Argues people are sensitive to ordering
  - Ordering can improve assessment

# Announcements

- D2 due next week!
- Presentations in place of regular material (come prepared to present!)
- Will send out scheduling poll for slot sign-ups.