Summarization: Overview

LING 573 — Systems & Applications
March 29, 2018

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Begin Recording!

Announcements

- Still 16 unassigned students!

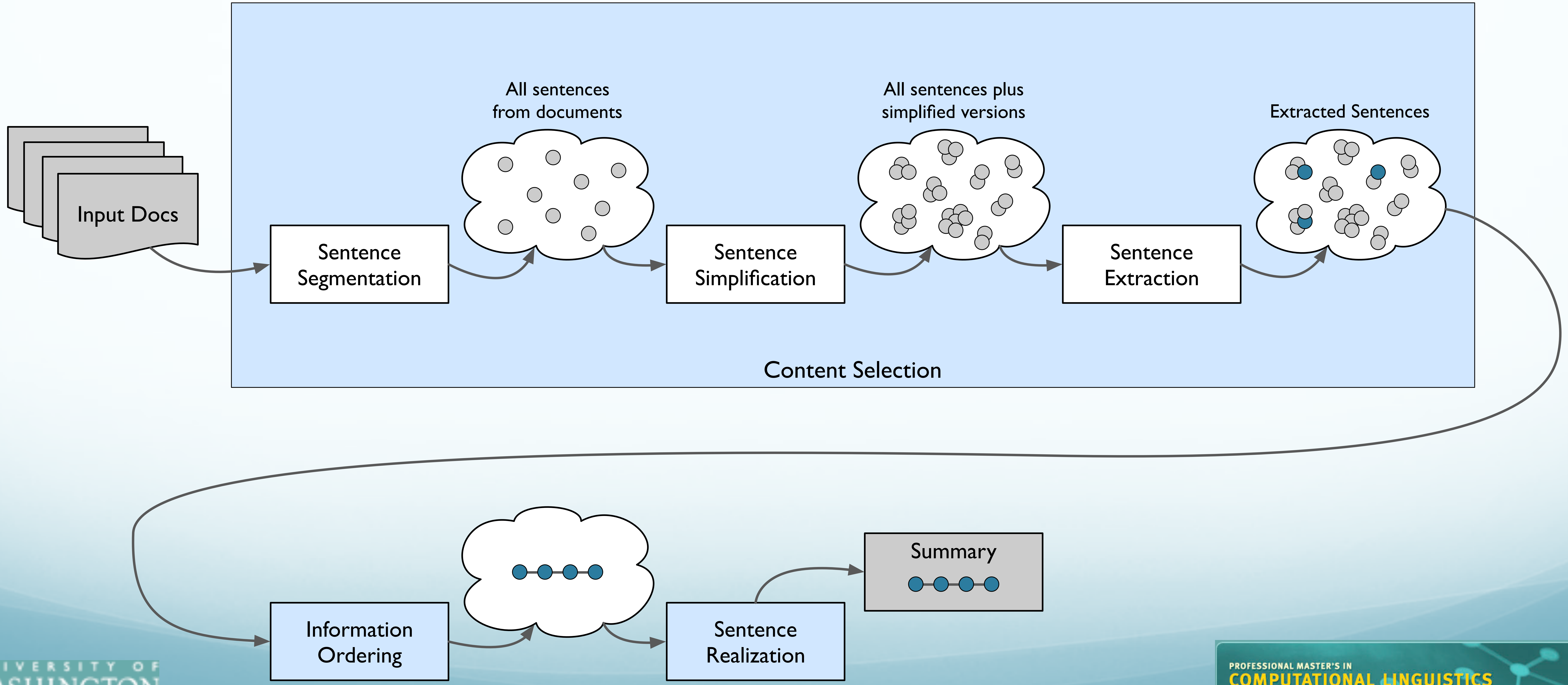
UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Roadmap

- Architecture of a Summarization System
- Summarization and Resources
- Evaluation
- Logistics Check-in, Deliverable #1

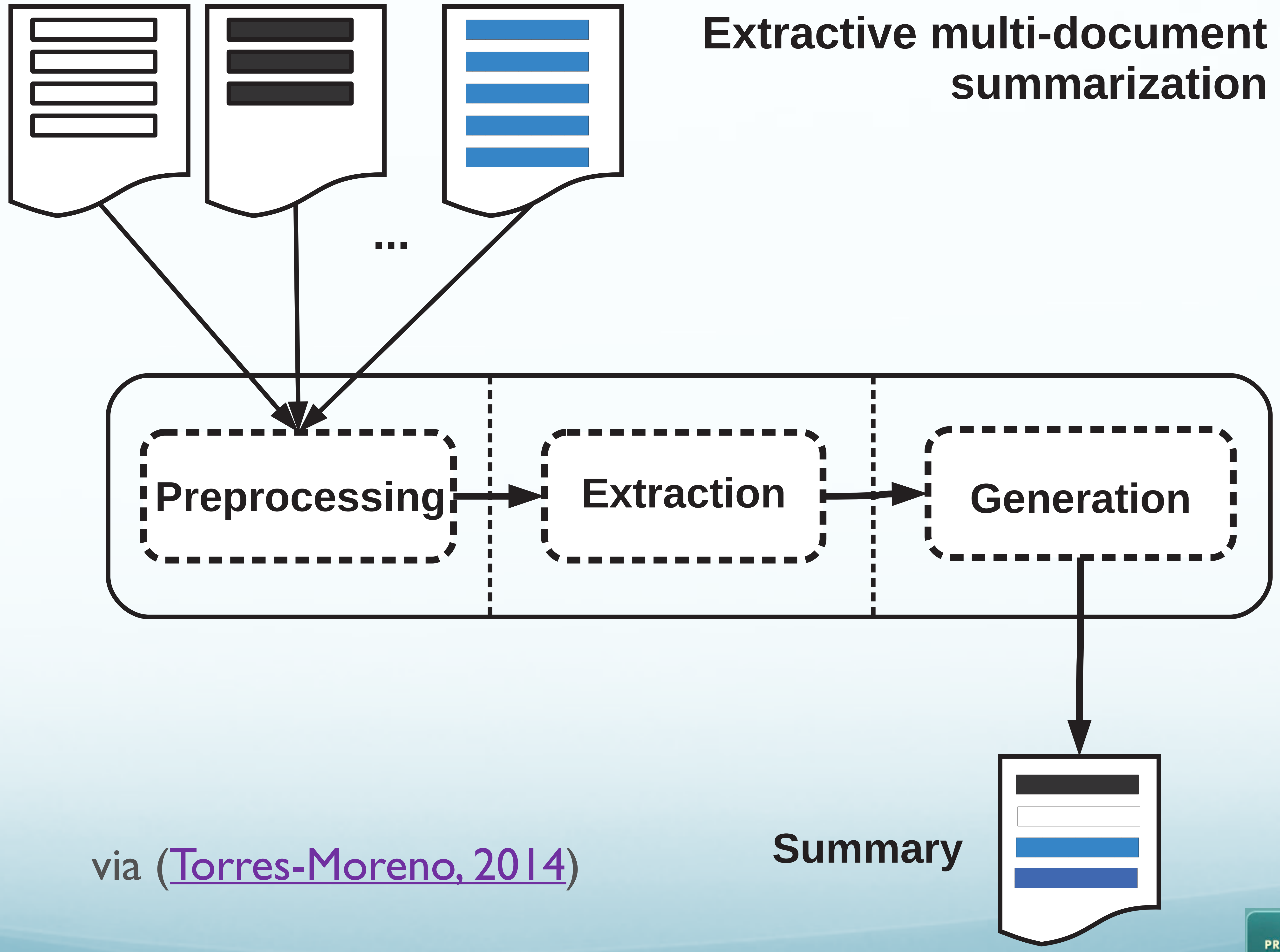
UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

General Architecture



UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

General Architecture



via (Torres-Moreno, 2014)

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

General Strategy

- Given a document (or set of documents)
  - Select the key content from the text
  - Determine the order to present that information
  - Perform cleanup or rephrasing to create coherent output
  - Evaluate the resulting summary

- Systems vary in structure, complexity, information

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

A Sample Extractive Approach

A Sample Extractive Approach

- For single document, extractive summarization

1. Segment the text into sentences (**Preprocessing**)
2. Identify the most prominent sentences (**Content Selection**)
3. Pick an order to present them (**Information Ordering**)
   - Order in which they appear?
   - Order of importance?
4. Do any necessary processing to improve coherence (**Sentence Realization**)
   - Shorten sentences, fix coreferences, etc.

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Content Selection

- Goal: Identify the most important/relevant information
- Commonly binary classification — **important** vs. **not important**
  - For each unit
    - e.g. sentence in the extractive case
  - Can be unsupervised or supervised

- *What makes a sentence extract-worthy?*

Cues to Saliency: Approaches

- **Word-based**
  - Compute a **topic signature** (Lin & Hovy, 2000, Radev et al, 2000)
  - Use distance/similarity metrics to compute unit relevancy to topic
- **Discourse-based**
  - Discourse saliency → Extract-worthiness
- **Multi-feature supervised**:
  - Cues include position, cue phrases, word salience

More Complex Settings

- **Multi-document case**
  - Key issue: redundancy
  - General idea — add salient content that is least similar to that already there
- **Topic/query-focused**
  - Ensure salient content related to topic/query
  - Prefer content more similar to topic
  - Alternatively, when given specific question types
    - Apply more Q/A information extraction oriented approach

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Information Ordering

- Goal — Determine presentation order for salient content
- Relatively trivial for single document extractive case
  - Just retain original document order of extracted sentences
- Multi-document case more challenging:
  - Story chronological order — insufficient alone
  - Discourse coherence and cohesion
    - Create discourse relations
    - Maintain cohesion among sentences, entities
- Template approaches also used with strong query

Content Realization

- Goal — Create a fluent, readable, compact output
- Abstractive approaches range from templates to full NLG
- Extractive approaches focus on
  - Sentence simplification/compression
    - Manipulation of parse tree to remove unneeded information
    - Both rule-based and machine-learned approaches
  - Reference presentation and ordering
    - Based on saliency hierarchy of mentions

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Example: Compression

- When it arrives sometime next year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.

Example: Compression

- ~~When it arrives sometime next year in new TV sets~~, **the V-chip will give parents** a ~~new and potentially revolutionary~~ **device to block out programs they don't want their children to see**.

Example: Coreference

Advisers do not blame **O'Neill**, but they recognize a shakeup would help indicate **Bush** was working to improve matters.

U.S. President George W. Bush pushed out **Treasury Secretary Paul O'Neill** and …

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Example: Coreference

Advisers do not blame **Treasury Secretary Paul O'Neill**, but they recognize a shakeup would help indicate **U.S. President George W. Bush** was working to improve matters.

Bush pushed out **O'Neill** and …

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Our Task

- TAC 2009/10/11 Shared Task
  - Multi-document summarization
  - Newswire Text
  - "**Guided**" (aka topic-oriented)
- **ROUGE** as primary evaluation metric.

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

System Resources

Resources: Training Data

- Sets of document(s) and summaries, info
  - Existing data sets from shared tasks
  - Manual summaries from other corpora
- Summary websites with pointers to source
- For technical domain, almost any paper
  - Articles require abstracts

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Resources: Components

- **Content Selection**:
  - Documents, corpora for term weighting
  - Sentence breakers
  - Semantic similarity tools (WordNet similarity)
  - Coreference resolver
  - Discourse parser
  - NER, IE
  - Topic segmentation
  - Alignment tools

Resources: Components

- **Information Ordering**:
  - Temporal processing
  - Coreference resolution
  - Lexical chains
  - Topic modeling
  - (Un)compressed sentence sets
- **Content realization**
  - Parsing
  - NP Chunking
  - Coreference

Summary Evaluation

Extrinsic Evaluation

- Given a task, does the summary enable users to perform the task?
  - As well as the full document set?
  - As fast or faster than the full document set?

- Hard to give a single, general approach, but some examples…

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Extrinsic Evaluation: Examples

- Time-limited fact-gathering
  - Answer questions about a topic/event
  - This can be a critical part of a job, e.g. a paralegal (or a NLP researcher!)
- Relevance assessment
- MOOC navigation
  - Raw video vs. auto-summary/index
  - Task completed faster w/summary (except expert MOOCers)

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Intrinsic Evaluation

- Need basic comparison to simple, naïve approach
- Baselines:
  - **Random**:
    - Select N random sentences
  - **Leading sentences**:
    - Select N leading sentences
    - For the news domain, surprisingly hard to beat
    - For reviews, last N sentences is best.

Intrinsic Evaluation

- Most common automatic method: **ROUGE**
  - "**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation"
  - Backronym inspired by BLEU for Machine Translation (MT)
  - Computes overlap between summaries
    - ROUGE-2 = Bigrams

$$ROUGE2 = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{bigram \in S} count_{match}(bigram)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{bigram \in S} count(bigram)}$$

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

ROUGE

- **Pros**:
  - Automatic evaluation allows tuning, given set of reference summaries
  - Simple measure
- **Cons**:
  - Even human summaries highly variable, disagreement
  - Poor handling of coherence
  - Okay for **extractive**, highly problematic for **abstractive**

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Pyramid Evaluation

- Content selection evaluation
  - Not focused on ordering, but rather, readability
- Aims to address issues in evaluation of summaries
  - Human variation
    - Significant disagreement, use multiple models
  - Analysis granularity
    - Not just "which sentence" — overlaps in sentence content
  - Semantic equivalence

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Pyramid Units

- **Step 1**: Extract Summary Content Units (SCUs)
  - Basic content meaning units
  - Semantic content
  - Roughly clausal
  - Identified manually by annotators from model summaries
  - Described in own words (possibly changing)

Example

- **A1**. The industrial espionage case …began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as a production director.
- **B3**. However, he left GM for VW under circumstances, which …were described by a German judge as "potentially the biggest-ever case of industrial espionage".
- **C6**. He left GM for VW in March 1993.
- **D6**. The issue stems from the alleged recruitment of GM's…procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues.
- **E1**. On March 16, 1993, … Agnacio Lopez De Arriortua, left his job as head of purchasing at General Motor's Opel, Germany, to become Volkswagen's Purchasing … director.
- **F3**. In March 1993, Lopez and seven other GM executives moved to VW overnight.

Example

- **A1**. The industrial espionage case …began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as a production director.
- **B3**. However, he left GM for VW under circumstances, which …were described by a German judge as "potentially the biggest-ever case of industrial espionage".
- **C6**. He left GM for VW in March 1993.
- **D6**. The issue stems from the alleged recruitment of GM's…procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues.
- **E1**. On March 16, 1993, … Agnacio Lopez De Arriortua, left his job as head of purchasing at General Motor's Opel, Germany, to become Volkswagen's Purchasing … director.
- **F3**. In March 1993, Lopez and seven other GM executives moved to VW overnight.

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Example SCUs

- **SCU1 (w=6)**: *Lopez left GM for VW*
  - **A1**. the hiring of Jose Ignacio Lopez, an employee of GM . . . by VW
  - **B3**. he left GM for VW
  - **C6**. He left GM for VW
  - **D6**. recruitment of GM's . . . Jose Ignacio Lopez
  - **E1**. Agnacio Lopez De Arriortua, left his job . . . at General Motor's Opel . . .to become Volkswagen's . . . Director
  - **F3**. Lopez . . . GM . . . moved to VW
- **SCU2 (w=3)**: *Lopez changes employers in March 1993*
  - **C6**. in March, 1993
  - **E1**. On March 16, 1993
  - **F3**. In March 1993

SCU:
(Weight = ?)

A. The cause of the fire was unknown.

B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.

C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.

D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

SCU: A Cable Car Caught Fire (Weight = 4)

A. The cause of the fire was unknown.

B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.

C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.

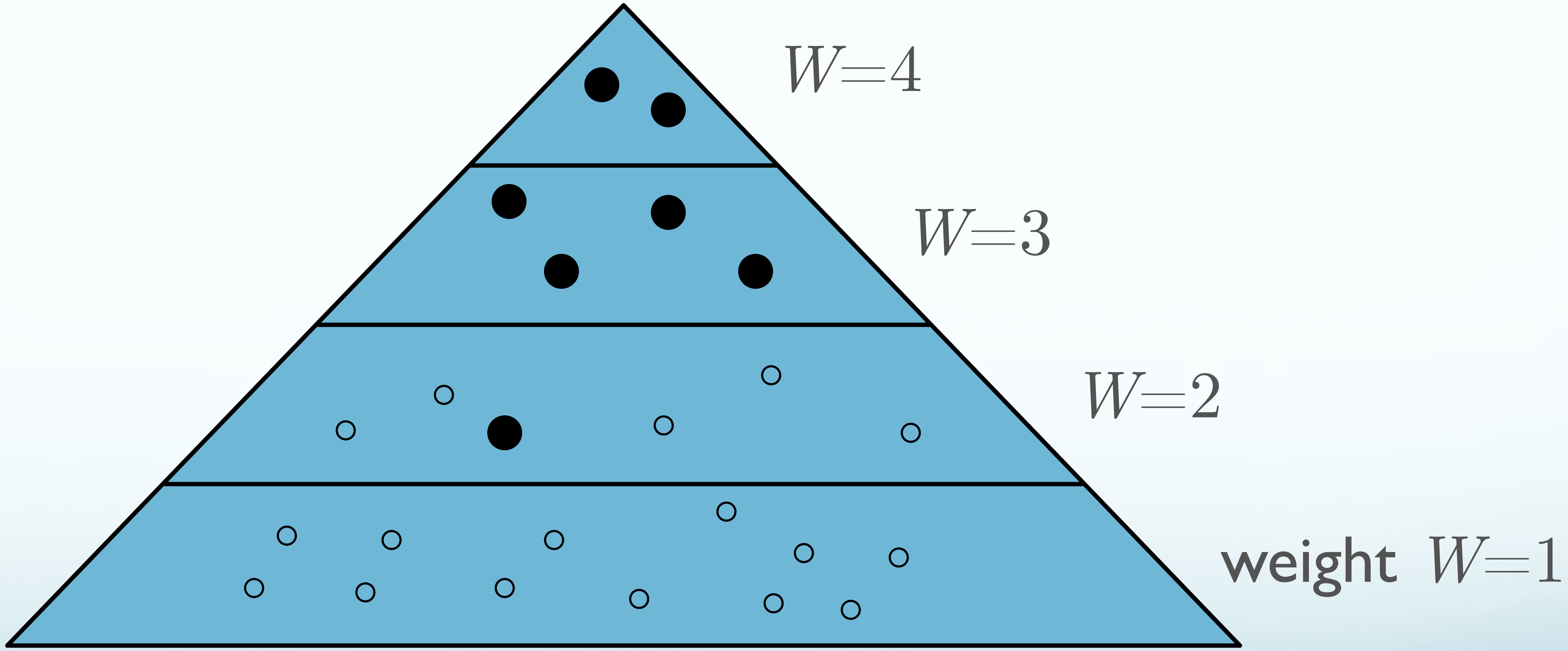
D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

Pyramid Building

- Step 2 — Scoring Summaries
  - Compute weights of SCUs
    - Weight = # of model summaries in which SCU appears
  - Create "pyramid"
    - n = maximum number of tiers in pyramid = # of model summaries
    - Actual # of tiers depends on degree of overlap
    - Highest tier: highest weight SCUs
      - Roughly Zipfian SCU distribution, so pyramidal shape
  - Optimal summary?
    - All from top tier, then all from top–1, until maximum summ. lenght reached.

Ideally Informative Summary

- Does not include any SCU from a lower tier unless all SCUs from higher tiers are included as well.



From [Nenkova et. al (2007)]

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Pyramid Scores

- T_i = tier with weight i SCUs
  - T_n = top tier; T_1 = bottom tier
  - D_i = # of SCUs in summary on T_i
  - Total weight of summary $D = \sum_{i=1}^{n} i * D_i$
  - Optimal score for X SCU summary $\sum_{i=j+1}^{n} i * |T_i| + j * (X - \sum_{i=j+1}^{n} |T_i|)$
    - (j is lowest tier in ideal summary)

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Pyramid Scores

- Original Pyramid Score:
  - Ratio of D to optimal possible score
    - Precision-oriented
- Modified Pyramid Score:
  - X_a = Average # of SCUs in model summaries
  - Ratio of D to Max (using X_a)
  - More recall oriented (most commonly used)

UNIVERSITY OF WASHINGTON
PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Correlation with Other Scores

Table VI. Pearson's Correlation Between the Different Evaluation Metrics Used in DUC 2005. Computed for 25 Automatic Peers Over 20 Test Sets.

	Pyr (mod)	Respons-1	Respons-2	ROUGE-2	ROUGE-SU4
Pyr (orig)	0.96	0.77	0.86	0.84	0.80
Pyr (mod)		0.81	0.90	0.90	0.86
Respons-1			0.83	0.92	0.92
Respons-2				0.88	0.87
ROUGE-2					0.98

- > 0.95: effectively indistinguishable
  - Two pyramid models, two ROUGE models
- **N.B.**: Two humans, only 0.83

Pyramid Model

- Pros:
  - Achieves goals of handling variation, abstraction, semantic equivalence
  - Can be done sufficiently reliably
  - Achieves good correlation with human assessors
- Cons:
  - Requires heavy manual annotation
    - Model summaries, also all system summaries
    - Content only

Deliverable #1

- Goals:
  - Set up for remainder of course
  - Form teams
  - Set up repository for version control
  - Create report outline
    - Using ACL style files
- Post to Group Roll-Call thread by the weekend

Preprocessing

Preprocessing Steps

- Segmentation — sentences, paragraphs, etc
- Tokenization — words, MWE
- Normalization — stemming
- Filtering — stopwords, punctuation (maybe)