

EECS 487: Introduction to Natural Language Processing

Instructor: Prof. Lu Wang

Computer Science and Engineering

University of Michigan

<https://web.eecs.umich.edu/~wangluxy/>

Logistics

- **Midterm:**
 - Time: Mar 17 (next Monday, no lecture), 5-7pm
 - Location: CHRYS220
 - Open book (can take any textbook, slides, cheat sheet, laptop, calculator; no internet access, no cellphone usage, can't communicate with others)
 - Topics: up to the lecture by Mar 10 (inclusive)
- **Schedule for rest of this semester** (also available on canvas):
 - Topics to be covered: summarization, question answering, sentiment analysis, machine, coreference, discourse, advanced topics of LLMs
 - Mar 31: meet with the instructor to discuss project progress (optional)
 - If you can't make it to the regular office hours, you're encouraged to sign up to chat about problems and challenges.
 - Project presentations are scheduled on Apr 9, 14, 16, & 21, team orderings and dates will be announced on piazza.

Outline

- ➔ • Text summarization tasks
 - Single vs. multiple documents, query-focused vs. generic, extractive vs. abstractive
- Summarization pipeline and content selection
 - How to detect salient sentences and salient words
- Summary evaluation
- Other aspects of summary quality
- Information ordering

Text Summarization

- **Goal:** produce an abridged version of a text that contains information that is important or relevant to a user.
- **Summarization Applications**
 - **outlines or abstracts** of any document, article, etc
 - **summaries** of email threads
 - **action items** from a meeting
 - **simplifying** text by compressing sentences



Speech Summarization

Phone Conversation



Lecture



Meeting



Talk Shows



Chat



Classroom



Broadcast News



Radio News

Why we need summarization?

- “Summaries as short as 17% of the full text length **speed up decision making twice**, with no significant degradation in accuracy.”
 - Does this document contain information that I am interested in?
 - Is this document worth reading?
- “**Query-focused summaries** enable users to find more relevant documents more accurately, with less need to consult the full text of the document.” [Mani et al., 2002]

What is the output

- Keywords
- Highlighted information in the input
- Chunks or speech directly from the input or paraphrase and aggregate the input in novel ways
- Modality: text, speech, video, graphics

What is the output

- Keywords
- Highlighted information in the input
- Chunks or speech directly from the input or paraphrase and aggregate the input in novel ways
- Modality: [text](#), speech, video, graphics

What to summarize (i.e. input)?

Single vs. multiple documents

- **Single-document summarization**

- Given a single document, produce
 - abstract (a paragraph)
 - outline (bullet points)
 - headline (one sentence)

- **Multiple-document summarization**

- Given a group of documents (usually relevant and pre-clustered), produce a gist of the content:
 - a series of news stories on the same event (this can be a timeline summarization)
 - a set of web pages about some topic or question (e.g. you want different perspectives on a certain policy or some medical treatment)

Example: Scientific article summarization

- Single-document summarization task:
 - Not only what the article is about, but also how it relates to work it cites → *summarize the article with regard to prior work*
 - *“the proposed method addresses the scalability issue...”*
- Multi-document summarization task:
 - Determine which approaches are criticized and which are supported → *summarize articles that cite a given article*
 - *“xx et al. presents an efficient algorithm...”*
 - *“Results by xx et al. shows...”*

Do we have a focus?

Query-focused Summarization vs. Generic Summarization

- **Generic summarization:**


- Summarize the content of one or multiple documents

- **Query-focused summarization:**

- Summarize one or multiple documents with respect to *an information need expressed in a user query*.
- a kind of complex **question answering**:
 - Answer a question by summarizing a document that has the information to construct the answer

Summarization for Question Answering or Search Engine: Featured Snippets

All Videos Images Shopping News Short videos Forums More Tools

 Search Labs | AI Overview


Generative AI (Gen AI) is a type of artificial intelligence that can create new content, such as text, images, videos, and music.

How it works

- Gen AI models are trained on large datasets.
- The models learn to recognize patterns in the data.
- Based on this learning, the models build predictive models.
- The models can then create new content based on prompts or inputs.

Use cases

- Create a short story based on a specific author's style
- Generate a realistic image of a person who doesn't exist
- Compose a symphony in the style of a famous composer
- Create a video clip from a textual description
- Generate personalized responses for customer service agents



Generative artificial intelligence - Wikipedia

Video generated by Sora with prompt Borneo wildlife on the Kinabatangan River. Generative AI trained on annotated vide...

Wikipedia

What is Generative AI? - Gen AI Explained - AWS

Generative artificial intelligence, also known as generative AI or gen AI for short, is a type of AI that can create new content a...

AWS

What is Generative AI? - Academic Excellence - Office of the Provost

What is Generative AI? * Generative artificial intelligence (Gen AI) is artificial intelligence that can generate text, images, o...


McMaster University

Show all

12

Summarization for Question Answering or Search Engine: Featured Snippets

[All](#) [Videos](#) [Images](#) [Shopping](#) [News](#) [Short videos](#) [Forums](#) [More](#) [Tools](#)

 Search Labs | AI Overview

Generative AI (Gen AI) is a type of artificial intelligence that can create new content, such as text, images, videos, and music.


How it works

- Gen AI models are trained on large datasets.
- The models learn to recognize patterns in the data.
- Based on this learning, the models build predictive models.
- The models can then create new content based on prompts or inputs.

Use cases

- Create a short story based on a specific author's style
- Generate a realistic image of a person who doesn't exist
- Compose a symphony in the style of a famous composer
- Create a video clip from a textual description
- Generate personalized responses for customer service agents

Considerations



Generative artificial intelligence - Wikipedia

Video generated by Sora with prompt Borneo wildlife on the Kinabatangan River. Generative AI trained on annotated vide...
Wikipedia

What is Generative AI? - Gen AI Explained - AWS

Generative artificial intelligence, also known as generative AI or gen AI for short, is a type of AI that can create new content a...
AWS

What is Generative AI? - Academic Excellence - Office of the Provost

What is Generative AI? * Generative artificial intelligence (Gen AI) is artificial intelligence that can generate text, images, o...
McMaster University

Show all

Looking at the output:

Extractive summarization vs. Abstractive summarization

- **Extractive summarization:**

- create the summary from phrases or sentences in the source document(s)
- will not use words that do not appear in the input




- **Abstractive summarization:**

- express the ideas in the source documents using (at least in part) *different words*


Extractive summarization vs. Abstractive summarization

what is generative ai

×



All Videos Images Shopping News Short videos Forums More Tools

 Search Labs | AI Overview

Learn more

Generative AI (Gen AI) is a type of artificial intelligence that can create new content, such as text, images, videos, and music.


How it works

- Gen AI models are trained on large datasets.
- The models learn to recognize patterns in the data.
- Based on this learning, the models build predictive models.
- The models can then create new content based on prompts or inputs.

Use cases

- Create a short story based on a specific author's style
- Generate a realistic image of a person who doesn't exist
- Compose a symphony in the style of a famous composer
- Create a video clip from a textual description
- Generate personalized responses for customer service agents

Considerations



Generative artificial intelligence - Wikipedia

Video generated by Sora with prompt Borneo wildlife on the Kinabatangan River. Generative AI trained on annotated vide...
Wikipedia

What is Generative AI? - Gen AI Explained - AWS

Generative artificial intelligence, also known as generative AI or gen AI for short, is a type of AI that can create new content a...
AWS

What is Generative AI? - Academic Excellence - Office of the Provost

What is Generative AI? * Generative artificial intelligence (Gen AI) is artificial intelligence that can generate text, images, o...
McMaster University

Show all

15

Extractive summarization

Sample article:

The Trump administration accused Russia on Thursday of engineering a series of cyberattacks that targeted American and European nuclear power plants and water and electric systems, and could have sabotaged or shut power plants off at will.

United States officials and private security firms saw the attacks as a signal by Moscow that it could disrupt the West's critical facilities in the event of a conflict.

They said the strikes accelerated in late 2015, at the same time the Russian interference in the American election was underway. The attackers had compromised some operators in North America and Europe by spring 2017, after President Trump was inaugurated.

Extractive summarization: sentence-level

Sample article:

The Trump administration accused Russia on Thursday of engineering a series of cyberattacks that targeted American and European nuclear power plants and water and electric systems, and could have sabotaged or shut power plants off at will.

United States officials and private security firms saw the attacks as a signal by Moscow that it could disrupt the West's critical facilities in the event of a conflict.

They said the strikes accelerated in late 2015, at the same time the Russian interference in the American election was underway. The attackers had compromised some operators in North America and Europe by spring 2017, after President Trump was inaugurated.

Extractive summarization: phrase-level

Sample article:

The **Trump administration** accused **Russia** on Thursday of engineering a series of **cyberattacks** that targeted **American and European nuclear power plants** and **water and electric systems**, and could have sabotaged or shut power plants off at will.

United States officials and private security firms saw the attacks as a signal by Moscow that it could disrupt the West's critical facilities in the event of a conflict.

They said the strikes accelerated in late 2015, at the same time the Russian interference in the American election was underway. The attackers had compromised some operators in North America and Europe by spring 2017, after President Trump was inaugurated.

Abstractive Summarization

- Input: Congratulations to Australia for seeing sense and dropping the ridiculous policy of not selecting their best players if they are playing overseas.
- Summary: Australia have seen sense by revamping their overseas selection policy.

Abstractive Summarization

- Input: Congratulations to Australia for seeing sense and dropping the ridiculous policy of not selecting their best players if they are playing overseas.
- Summary: Australia have seen sense by revamping their overseas selection policy.
- How does a model achieve this?
- Large language models, generative AI

Most efficient and generalizable systems

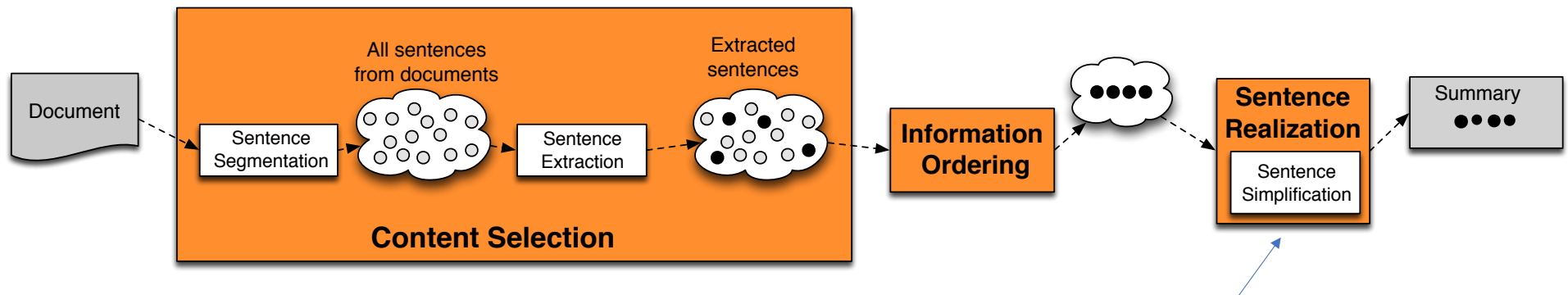
- Use shallow analysis methods (frequent words)
- Work by sentence selection
 - Identify important sentences and piece them together to form a summary

Outline

- Text summarization tasks
 - Single vs. multiple documents, query-focused vs. generic, extractive vs. abstractive
- ➔ • Summarization pipeline and content selection
 - How to detect salient sentences and salient words
- Summary evaluation
- Other aspects of summary quality
- Information ordering

Summarization: Three Stages

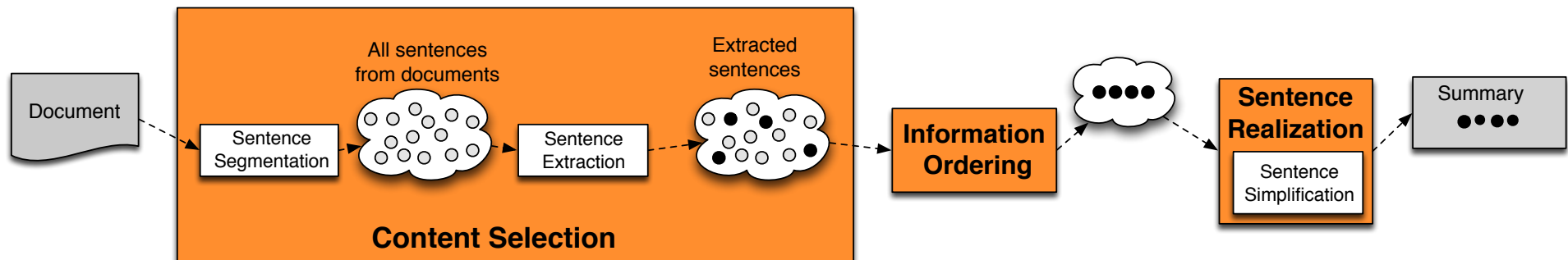
1. **content selection**: choose sentences to extract from the document
2. **information ordering**: choose an order to place them in the summary
3. **sentence realization**: clean up the sentences (e.g. removing redundancy)



Other operations: *sentence fusion* (multiple sentences are transformed into one sentence), *compression* (longer sentences are transformed into shorter ones), etc.

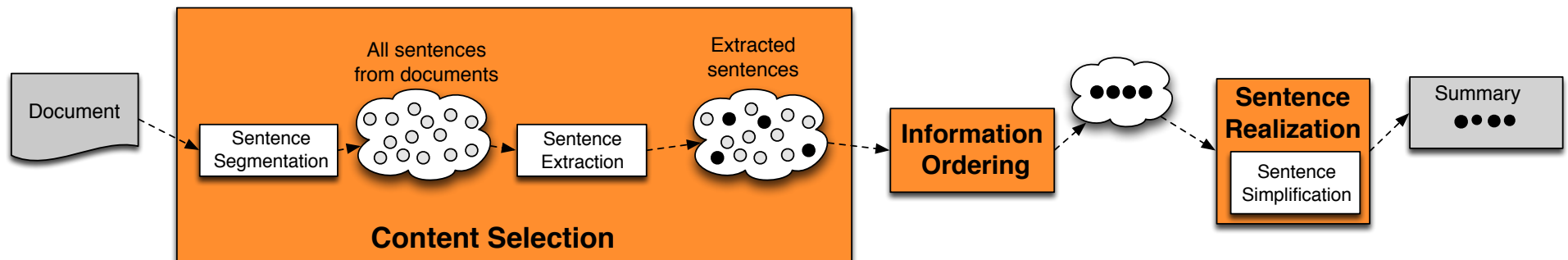
Basic Summarization Algorithm (extractive)

1. content selection: choose sentences to extract from the document
2. information ordering: just use document order
3. sentence realization: keep original sentences



Basic Summarization Algorithm (extractive)

1. **content selection**: choose sentences to extract from the document
2. **information ordering**: just use document order
3. **sentence realization**: keep original sentences



Unsupervised content selection

- What words or sentences can be considered as important?

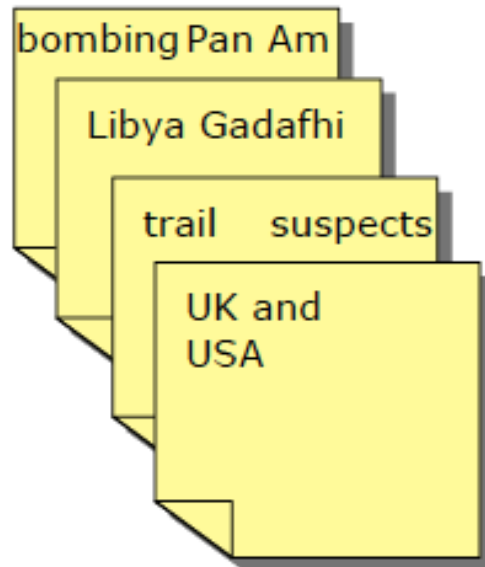
Frequency as document topic proxy

- Simple intuition, look only at the document(s)
 - Words that repeatedly appear in the document are likely to be related to the topic of the document (single document)
 - Sentences that repeatedly appear in different input documents represent themes in the input (multiple documents)
- But what appears in **other documents** is also helpful in determining the topic
 - Background corpus probabilities/weights for word

What is an article about?

- Word probability/frequency
 - Frequent content words would be indicative of the topic of the article
- In multi-document summarization, words or facts repeated in the input are more likely to appear in human summaries [Nenkova et al., 2006]

INPUT



WORD PROBABILITY TABLE

Word	Probability
pan	0.0798
am	0.0825
libya	0.0096
suspects	0.0341
gadafhi	0.0911
trail	0.0002
....	
usa	0.0007

SUMMARY

Libya refuses
to surrender
two Pan Am
bombing
suspects



Main steps in sentence selection according to word probabilities

- Step 1: estimate word weights (probabilities)
- Step 2: estimate sentence weights
- Step 3: choose best sentence
- Step 4: update word weights
- Step 5: go to step 2 if length not reached

Main steps in sentence selection according to word probabilities

- Step 1: estimate word weights (probabilities)
- Step 2: estimate sentence weights (how?)
- Step 3: choose best sentence
- Step 4: update word weights (why?)
- Step 5: go to step 2 if length not reached

Main steps in sentence selection according to word probabilities

- Step 1: estimate word weights (probabilities)
- Step 2: estimate sentence weights
- Step 3: choose best sentence
- Step 4: update word weights
- Step 5: go to step 2 if length not reached

Our
focus

- Select highest scoring sentence

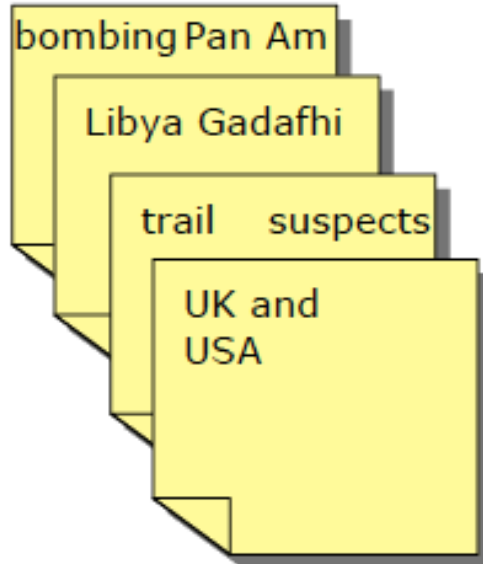
$$Score(S) = \frac{1}{|S|} \sum_{w \in S} p(w)$$

- Update word probabilities for the selected sentence to reduce redundancy

$$p^{new}(w) = p^{old}(w).p^{old}(w)$$

- Repeat until desired summary length

INPUT



WORD PROBABILITY TABLE

Word	Probability
pan	0.0798
am	0.0825
libya	0.0096
suspects	0.0341
gadafhi	0.0911
trail	0.0002
....	
usa	0.0007

SUMMARY

Libya refuses
to surrender
two Pan Am
bombing
suspects



Obvious shortcomings of the pure frequency approaches

- Does not take account of paraphrases or related words
 - bombing -- explosion
 - suspects -- trail
 - Gadhafi -- Libya
- Does not take into account evidence from other documents
 - Function words: prepositions, articles, etc.
 - Domain words: “cell” in cell biology articles
- Does not take into account many other aspects (relations, events, etc)!
 - Semantic in general!

Salient words

- Intuition dating back to Luhn (1958):
 - Choose sentences that have **salient** or **informative** words
- Two approaches to defining salient words
 1. **tf-idf**: weigh each word w_i in document j by tf-idf
$$weight(w_i) = tf_{ij} \times idf_i$$
 2. **topic signature**: choose a smaller set of salient words
 - log-likelihood ratio (LLR) test

Topic words (or topic signatures)

- Which words in the input are most descriptive?
- Instead of assigning probabilities or weights to all words, divide words into **two classes: descriptive or not**
- For iterative sentence selection approach, **the binary distinction** is key to the advantage over frequency and TF*IDF

Example input and associated topic words

- Input for summarization: articles relevant to the following user need

Title: Human Toll of Tropical

Storms Narrative: What has been the human toll in death or injury of tropical storms in recent years? Where and when have each of the storms caused human casualties? What are the approximate total number of casualties attributed to each of the storms?

Topic Words

ahmed, allison, andrew, bahamas, bangladesh, bn, caribbean, carolina, caused, cent, coast, coastal, croix, cyclone, damage, destroyed, devastated, disaster, dollars, drowned, flood, flooded, flooding, floods, florida, gulf, ham, hit, homeless, homes, hugo, hurricane, insurance, insurers, island, islands, lloyd, losses, louisiana, manila, miles, nicaragua, north, port, pounds, rain, rains, rebuild, rebuilding, relief, remnants, residents, roared, salt, st, storm, storms, supplies, tourists, trees, tropical, typhoon, virgin, volunteers, weather, west, winds, yesterday.

Formalizing the problem of identifying topic words

- Given
 - t : a word that appears in the input
 - T : cluster of articles on a given topic (input)
 - NT : articles not on topic T (background corpus)
- Decide if t is a topic word or not
- Words that have (almost) the same probability in T and NT are not topic words

H1: $P(t|T) = P(t|NT) = p$ (t is not a descriptive term for the topic)

H2: $P(t|T) = p_1$ and $P(t|NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

Computing probabilities

- View a text (i.e. a sequence of words) as a sequence of Bernoulli trials
 - A word is either our term of interest t or not
 - The likelihood of observing term t which occurs with probability p in a text consisting of N words is given by

$$b(k, N, p) = \binom{N}{k} p^k (1 - p)^{N-k}$$

Testing which hypothesis is more likely: log-likelihood ratio test

H1: $P(t|T) = P(t|NT) = p$ (t is not a descriptive term for the topic)

H2: $P(t|T) = p_1$ and $P(t|NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

$$\lambda = \frac{\text{Likelihood of the data given H1}}{\text{Likelihood of the data given H2}}$$

Testing which hypothesis is more likely: log-likelihood ratio test

H1: $P(t|T) = P(t|NT) = p$ (t is not a descriptive term for the topic)

H2: $P(t|T) = p_1$ and $P(t|NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

$$\lambda = \frac{\text{Likelihood of the data given H1}}{\text{Likelihood of the data given H2}}$$

$-2 \log \lambda$ has a known statistical distribution: chi-square

At a given significance level, we can decide if a word is descriptive of the input or not.

More information can be found: https://en.wikipedia.org/wiki/Likelihood-ratio_test (not required for this course)

Unsupervised content selection

- Topic signatures are assigned with weight of 1

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log \lambda(w_i) > 10 \\ 0 & \text{otherwise} \end{cases} \quad \text{significance level at 0.001}$$

Topic signature-based content selection with queries

- choose words that are informative either
 - by log-likelihood ratio (LLR) test
 - or by appearing in the query (if there is question)

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log \lambda(w_i) > 10 \\ 1 & \text{if } w_i \in \text{question} \\ 0 & \text{otherwise} \end{cases}$$

(could learn more complex weights)

- Weigh a sentence (or window) by weight of its words:

$$weight(s) = \frac{1}{|S|} \sum_{w \in S} weight(w)$$

So far, it's unsupervised content selection?

How to do it with supervision?

- Given:
 - a labeled training set of good summaries for each document
- Align:
 - the sentences in the document with sentences in the summary
 - Or ask human to select sentences
- Extract features
 - position (first K sentence?)
 - length of sentence
 - word informativeness, cue phrases
- Train
 - a binary classifier (put sentence in summary? yes or no)


Problems:

- hard to get labeled training data (sometimes only abstractive summaries are available)
- alignment difficult
- even the same person would select different sentences if she performs the task at different times
- performance not better than unsupervised algorithms

So in practice:

- **Unsupervised content selection is more common**

Outline

- Text summarization tasks
 - Single vs. multiple documents, query-focused vs. generic, extractive vs. abstractive
- Summarization pipeline and content selection
 - How to detect salient sentences and salient words
-  • Summary evaluation
- Other aspects of summary quality
- Information ordering

Evaluating Summaries: ROUGE

Human 1: Water spinach is a green leafy vegetable grown in the tropics.

Human 2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Human 3: Water spinach is a commonly eaten leaf vegetable of Asia.

- System answer: Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Lin and Hovy 2003

- Intrinsic metric for automatically evaluating summaries
 - Not as good as human evaluation (e.g. “Did this answer the user’s question?”)
 - But much more convenient, and still used nowadays!
- Given a document D, and an automatic summary X:
 1. Have N humans produce a set of reference summaries of D
 2. Run system, giving automatic summary X
 3. What percentage of the bigrams from the reference summaries appear in X?

$$ROUGE - 2 = \frac{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \min(\text{count}(i, X), \text{count}(i, S))}{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \text{count}(i, S)}$$

A ROUGE example:

Q: “What is water spinach?”

$$ROUGE-2 = \frac{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \min(\text{count}(i, X), \text{count}(i, S))}{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \text{count}(i, S)}$$

Human 1: Water spinach is a green leafy vegetable grown in the tropics.


Human 2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Human 3: Water spinach is a commonly eaten leaf vegetable of Asia.

- System answer: Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

- $ROUGE-2 = \frac{3 + 3 + 6}{10 + 10 + 9} = 12/29 = .43$

Outline

- Text summarization tasks
 - Single vs. multiple documents, query-focused vs. generic, extractive vs. abstractive
- Summarization pipeline and content selection
 - How to detect salient sentences and salient words
- Summary evaluation
-  • Other aspects of summary quality
- Information ordering

How to measure redundancy?

Author JK Rowling has won her legal battle in a New York court to get an unofficial Harry Potter encyclopedia banned from publication.

A U.S. federal judge in Manhattan has sided with author J.K. Rowling and ruled against the publication of a Harry Potter encyclopedia created by a fan of the book series.

How to measure redundancy?

Author JK Rowling has won her legal battle in a New York court to get an unofficial Harry Potter encyclopedia banned from publication.

A U.S. federal judge in Manhattan has sided with author J.K. Rowling and ruled against the publication of a Harry Potter encyclopedia created by a fan of the book series.

Sample features:

- Frequency of unigrams or bigrams
- Word overlaps between sentences
 - Further consider synonyms (e.g. using WordNet)
- Embedding-based text similarity between sentences

How to measure fluency (or coherence)?

- A fluent sentence:

The new legal classification will entitle the workers to more pay and benefits.

How to measure fluency (or coherence)?

- A fluent sentence:

The new legal classification will entitle the workers to more pay and benefits.

- Less fluent sentences:

The the new legal classification will entitle will entitle the workers to more pay and benefits.

- Sample features: repetition of unigrams and bigrams

How to measure fluency (or coherence)?

- A fluent sentence:

The new legal classification will entitle the workers to more pay and benefits.

- Less fluent sentences:


The the new legal classification will entitle will entitle the workers to more pay and benefits.

- Sample features: repetition of unigrams and bigrams

The new legal classification **will the workers** to more pay and benefits.

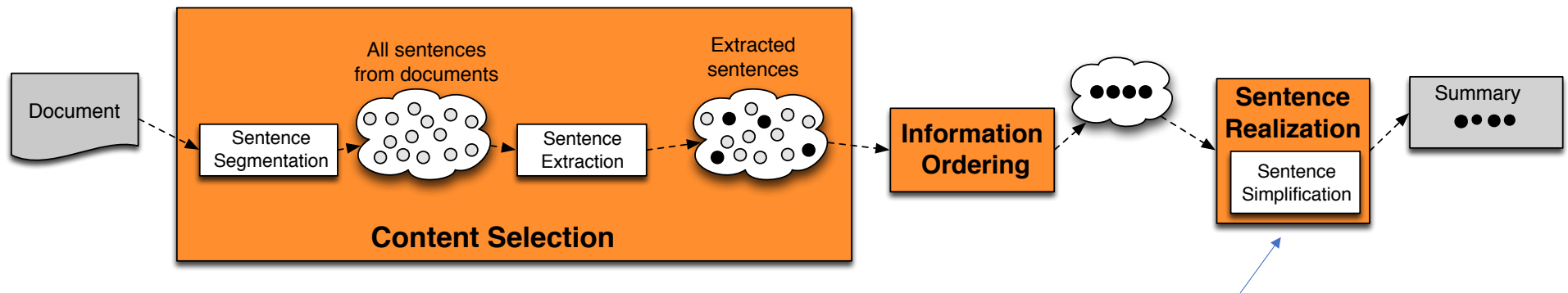
- Sample features: missing subject, object, or main verb?

Outline

- Text summarization tasks
 - Single vs. multiple documents, query-focused vs. generic, extractive vs. abstractive
- Summarization pipeline and content selection
 - How to detect salient sentences and salient words
- Summary evaluation
- Other aspects of summary quality
-  • Information ordering

Summarization: Three Stages

1. content selection: choose sentences to extract from the document
2. **information ordering**: choose an order to place them in the summary
3. sentence realization: clean up the sentences (e.g. removing redundancy)



Other operations: *sentence fusion* (multiple sentences are transformed into one sentence), *compression* (longer sentences are transformed into shorter ones), etc

Information Ordering

- In what order to present the selected sentences?
 - An article with permuted sentences will not be easy to understand
- Very important for multi-document summarization
 - Sentences coming from different documents

Information Ordering

- **Chronological ordering:**
 - Order sentences by the date of the document (for summarizing news) (Barzilay, Elhadad, and McKeown 2002)
- **Coherence:**
 - Choose orderings that make neighboring sentences similar (by cosine).
 - Choose orderings in which neighboring sentences discuss the same entity (Barzilay and Lapata 2007)
- **Topical ordering**
 - Learn the ordering of topics in the source documents

Automatic summary edits: advanced topics

- Some expressions might not be appropriate in the new context
 - References:
 - he
 - Putin
 - Russian Prime Minister Vladimir Putin
 - Discourse connectives
 - However, moreover, subsequently

Before and After

Pinochet was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. **Pinochet** has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. **Pinochet** was detained in the London clinic while recovering from back surgery.

Gen. Augusto Pinochet, the former Chilean dictator, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. **Pinochet** has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. **Pinochet** was detained in the London clinic while recovering from back surgery.