CS 6120/CS4120: Natural Language Processing

Instructor: Prof. Lu Wang
College of Computer and Information Science
Northeastern University
Webpage: www.ccs.neu.edu/home/luwang

Logistics

- I have office hour after today’s class at 258 WVH, 5:15pm-6:15pm.

- Final exam: 4/25, 8am-10am, 108 WVH (this room!)
  - Open book
  - You can bring in any notes and/or textbooks.
  - You can use your computer (e.g. laptops), but no access to Internet.
  - No cell phone! No messenger!

- We will have office hours until 4/24!

Schedule

- Topics to be covered: sentiment analysis, opinion mining, NLP for social media, dialogue systems and chatbots

- April 3: project feedback! **3:25-6:15pm in 258 WVH**.
  - Will create a thread on piazza
  - Teams that present on April 6 will have priority to pick time slots!

- April 6, 10, 13: 21 projects presentation!

Progress report content (also on Piazza!)

- Deadline extended to 3/26, 11:59pm
- What changes you have made for the task compared to the proposal, including problem/task, models, datasets, or evaluation methods? If there is any change, please explain why.
- Describe data preprocessing process. This includes data cleaning, selection, feature generation or other representation you have used, etc.
- What methods or models you have tried towards the project goal? And why do you choose the methods (you can including related work on similar task or relevant tasks)?
- What results you have achieved up to now based on your proposed evaluation methods? What is working or What is wrong with the model?
- How can you improve your models? What are the next steps?

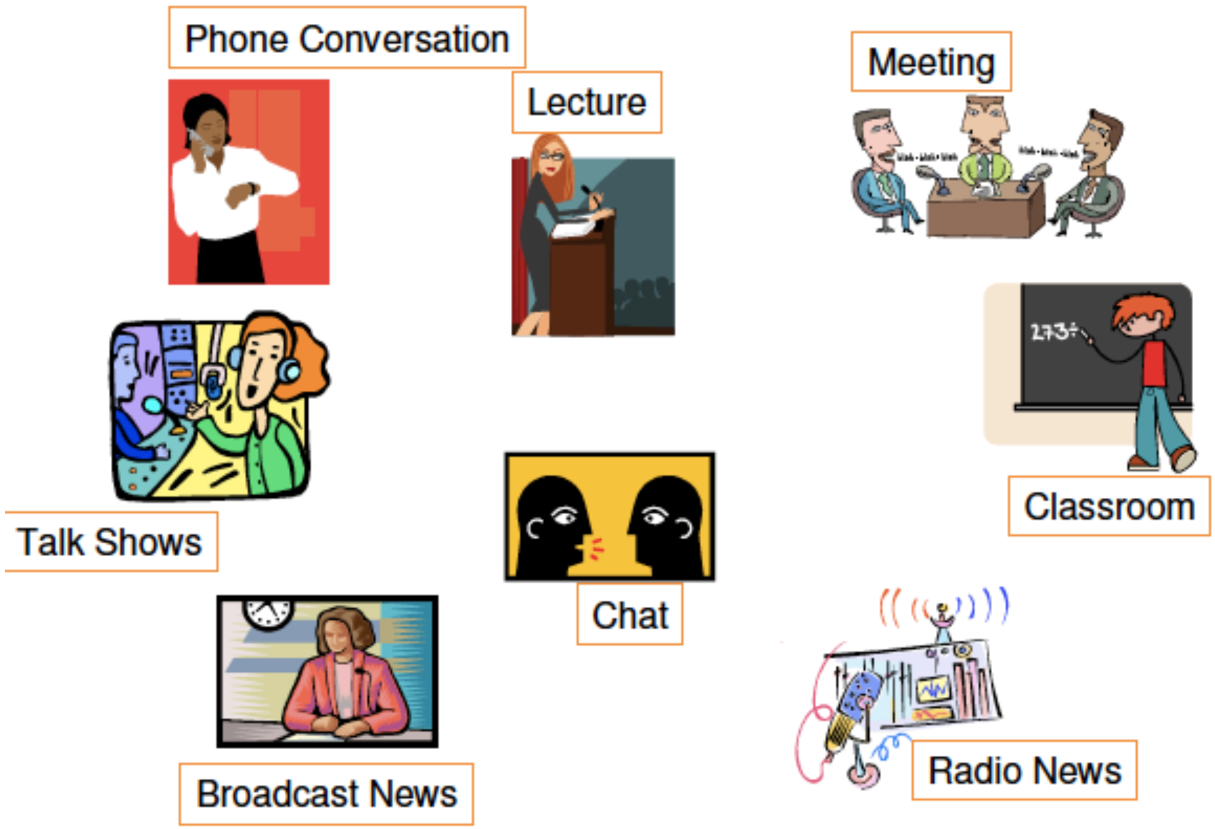
- Length: Length: 2 page (or more if necessary). Single space if MS word is used. Or you can choose latex templates, e.g. https://www.acm.org/publications/proceedings-template or http://icml.cc/2015/?page_id=151

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



- “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 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]
- One of our research projects at Northeastern: help quickly diagnose information retrieval (IR) systems -> is the system working as what we want?

Example

- Query: “Cyberattacks by Russian”
- Summary for returned doc 1:
    - *Over the last year, Russian hackers have gone from infiltrating business networks of energy, water and nuclear plants to worming their way into control rooms.*
- Summary for returned doc 2:
    - *The UK, US, France and Germany say there is no plausible alternative explanation to Russian responsibility.*

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?
Single vs. multiple documents

- **Single-document summarization**
  - Given a single document, produce
    - abstract
    - outline
    - headline
- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content:
    - a series of news stories on the same event
    - a set of web pages about some topic or question

Multi-document summarization

- Very useful for presenting and organizing search results
  - Many results are very similar, and grouping closely related documents helps cover more event facets
  - Summarizing similarities and differences between documents

Scientific article summarization

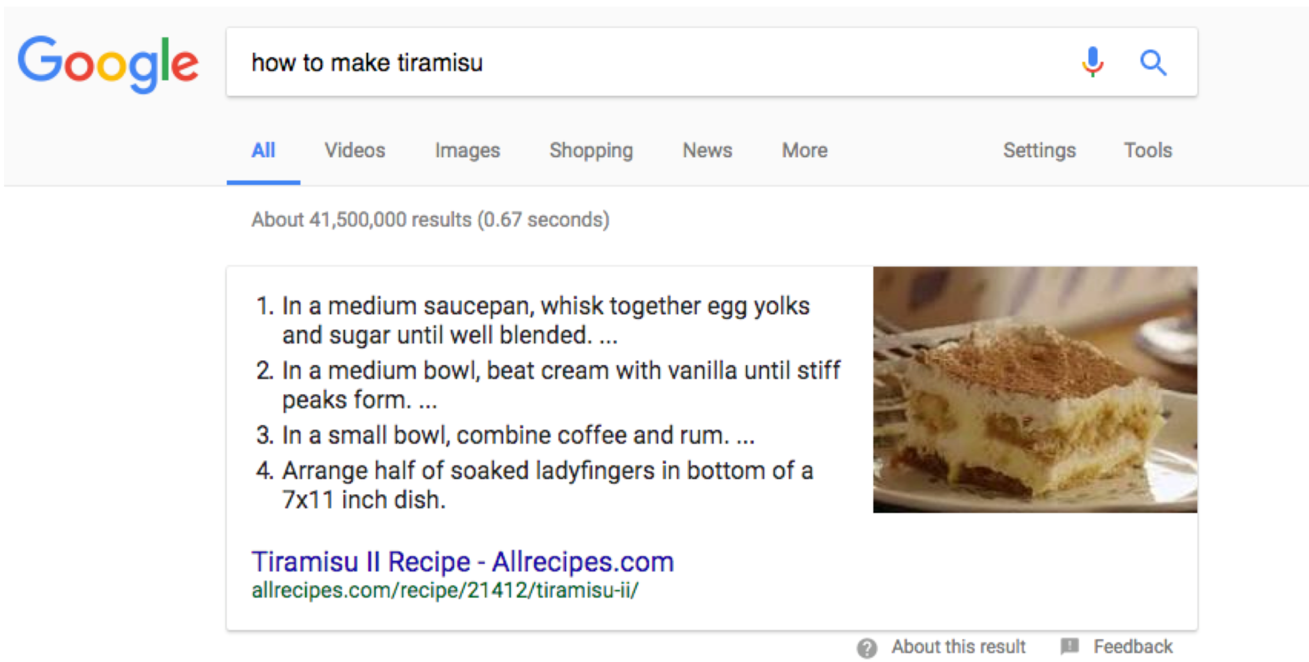
- Not only what the article is about, but also how it relates to work it cites
- Determine which approaches are criticized and which are supported
  - Automatic genre specific summaries are more useful than original paper abstracts

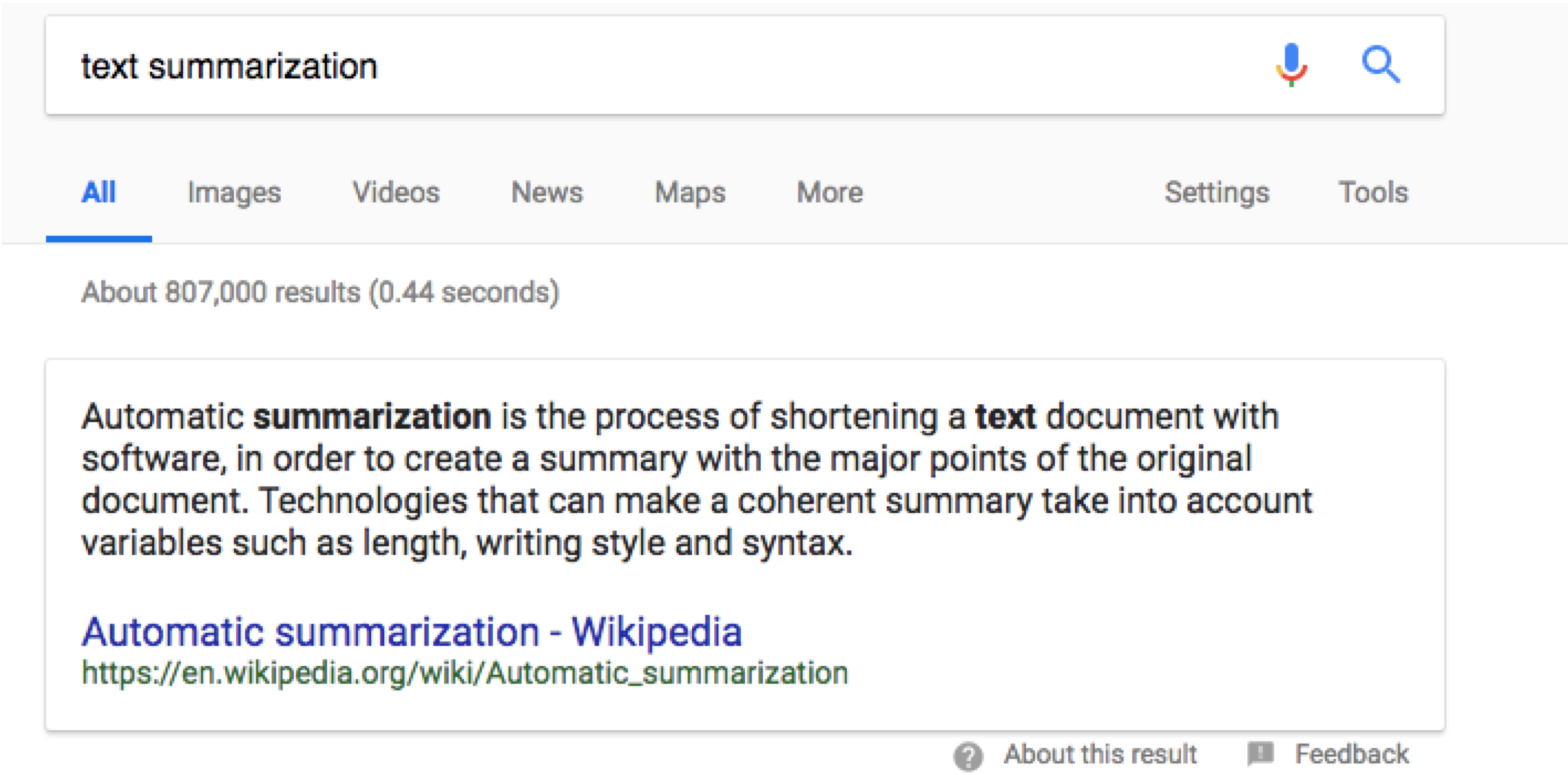
Query-focused Summarization & Generic Summarization

- Generic summarization:
  - Summarize the content of a document
- Query-focused summarization:
  - Summarize a document 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: Featured Snippets

- **Create** **snippets** summarizing a web page for a query (could be paragraphs)




text summarization
All
Images
Videos
News
Maps
More
Settings
Tools
About 807,000 results (0.44 seconds)
Automatic **summarization** is the process of shortening a **text** document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax.
Automatic summarization - Wikipedia
https://en.wikipedia.org/wiki/Automatic_summarization
About this result
Feedback

Summarization for Question Answering: Multiple documents

Create **answers** to complex questions summarizing multiple documents.

- Instead of giving a snippet for each document
- Create a cohesive answer that combines information from each document

Extractive summarization & Abstractive summarization

- Extractive summarization:
  - create the summary from phrases or sentences in the source document(s)
- Abstractive summarization:
  - express the ideas in the source documents using (at least in part) different words

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

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

Sample article:

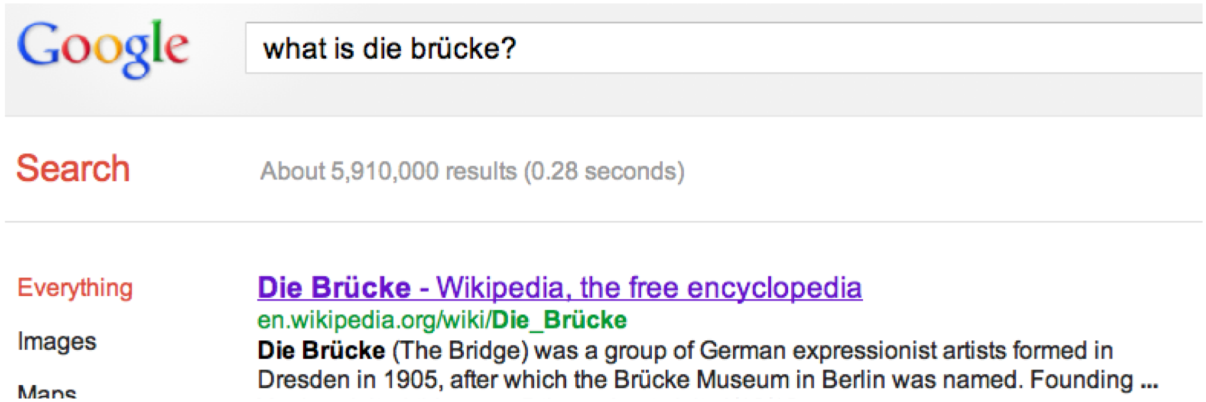
The Trump administration accused Russia on Thursday of engineering a series of cyberattacks that targeted American and European nuclear power plants and water and electric systems, and could have sabotaged or shut power plants off at will.

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.

- Word statistics
- Cue phrases
- Section headers (e.g. introduction in scientific articles)
- Sentence position

Simple baseline (extractive summarization): take the first sentence



Google

what is die brücke?

Search

About 5,910,000 results (0.28 seconds)

Everything

Images

Maps

Die Brücke - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/**Die_Brücke**

Die Brücke (The Bridge) was a group of German expressionist artists formed in Dresden in 1905, after which the Brücke Museum in Berlin was named. Founding ...

Die Brücke

From Wikipedia, the free encyclopedia

For other uses, see Die Brücke (disambiguation).

Die Brücke (**The Bridge**) was a group of German expressionist artists formed in Dresden in 1905, after which the **Brücke Museum in Berlin** was named. Founding members were Fritz Bleyl, Erich Heckel, Ernst Ludwig Kirchner and Karl Schmidt-Rottluff. Later members were Emil Nolde, Max Pechstein and Otto Mueller. The seminal group had a major impact on the evolution of modern art in the 20th century and the creation of expressionism.[1]

Die Brücke is sometimes compared to the Fauves. Both movements shared interests in primitivist art. Both

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.

Most current systems

- Use shallow analysis methods
  - Rather than full understanding
- Mostly extractive!---disadvantages?
  - Work by sentence selection
  - Identify important sentences and piece them together to form a summary

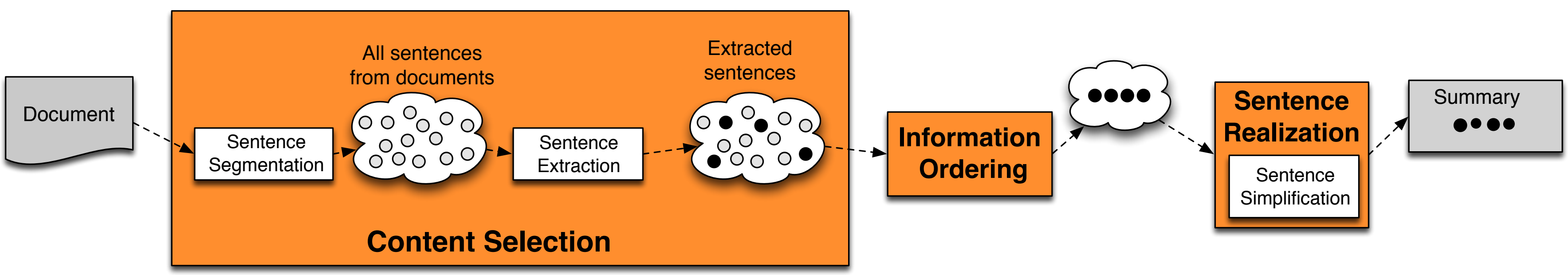
Knowledge-based and Advanced Systems

- Use more sophisticated natural language processing

- Discourse information
  - Resolve anaphora, text structure
- Use external lexical resources
  - Wordnet, adjective polarity lists, opinion
- Using machine learning models
- Towards abstractive summarization

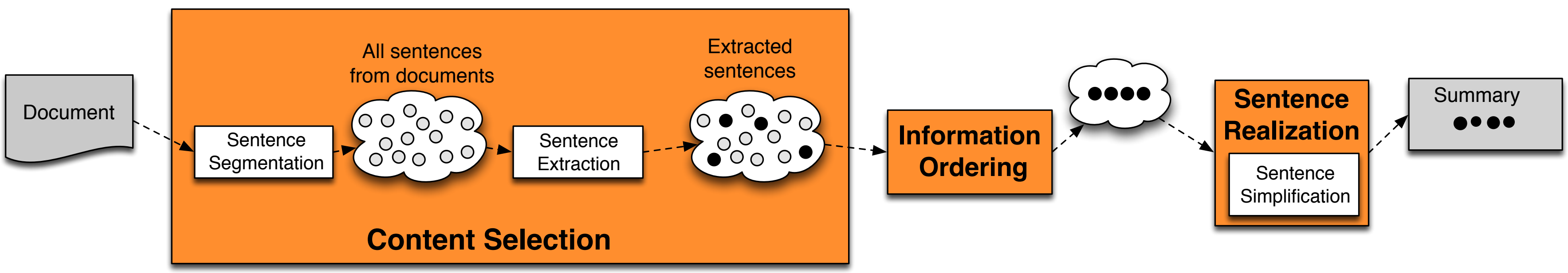
Summarization: Three Stages

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



Basic Summarization Algorithm

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



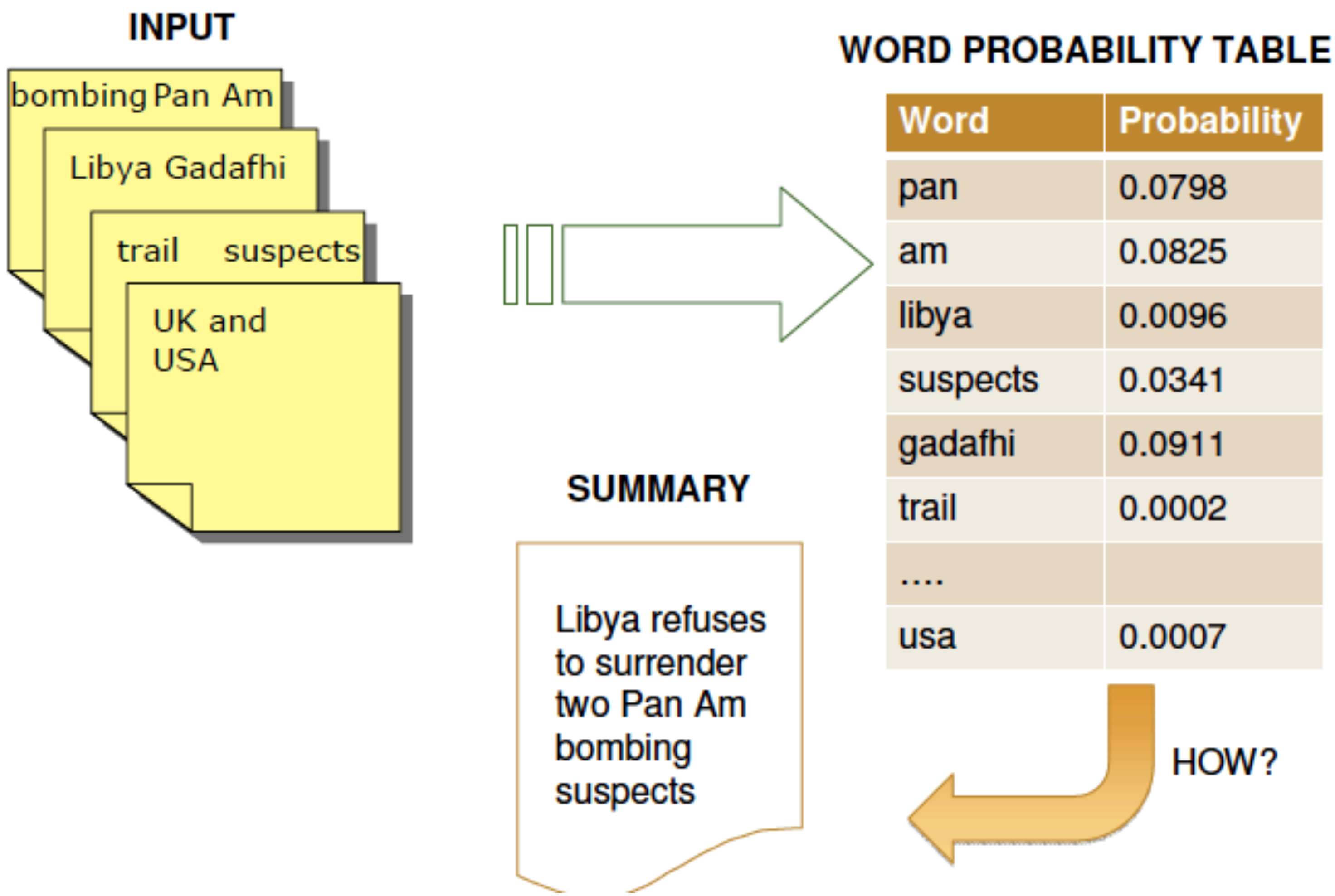
Unsupervised content selection

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
  - Sentences that repeatedly appear in different input documents represent themes in the input
- 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
  - Proposed by Luhn in 1958 [Luhn 1958]
  - 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
bombing Pan Am
Libya Gadafhi
trail suspects
UK and
USA
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
HOW?
SUMMARY
Libya refuses
to surrender
two Pan Am
bombing
suspects

How to deal with 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.

Main steps in sentence selection according to word probabilities

Step 1 Estimate word weights (probabilities)

Step 2 Estimate sentence weights

$$Weight(Sent) = CF(w_i \in Sent)$$

Step 3 Choose best sentence

Step 4 Update word weights

Step 5 Go to 2 if desired length not reached

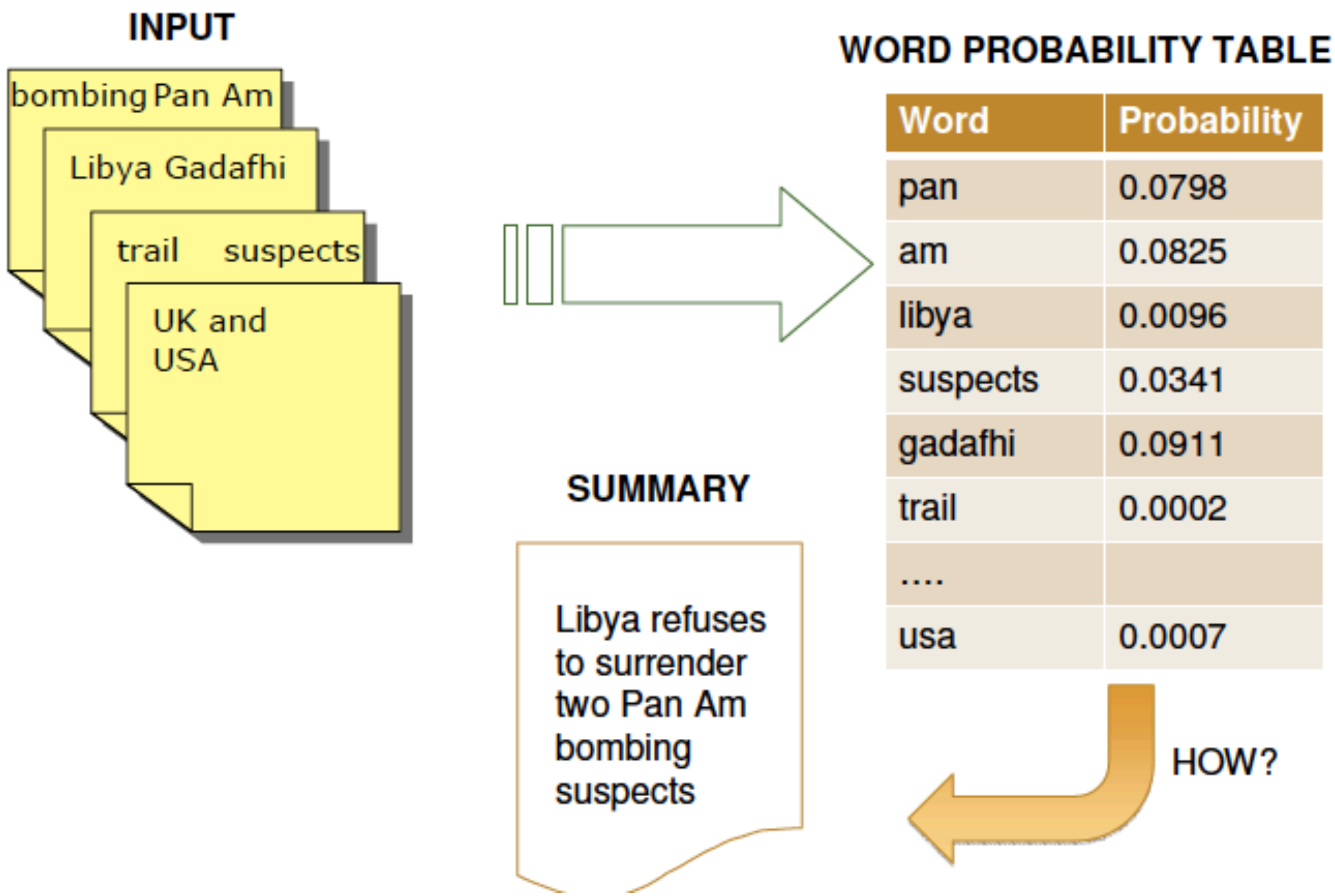
- Select highest scoring sentence

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

- Update word probabilities for the selected sentence to reduce redundancy

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

- Repeat until desired summary length


INPUT
bombing Pan Am
Libya Gadafhi
trail suspects
UK and USA
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
HOW?
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
  - 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!
  - Semantic in general!

Topic words (topic signatures)

H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 2:2, 159-165.

- 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
     - mutual information
     - log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

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

Topic words (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
- Systems based on topic words have proven to be the most successful in official summarization evaluations

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 as a sequence of Bernoulli trails
  - 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

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

- Estimate the probability of t in three ways
  - Input + background corpus combines
  - Input only
  - Background only

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.

This feature is used in the best performing systems for multi-document summarization of news [Lin and Hovy, 2000; Conroy et al., 2006]

Unsupervised content selection

H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 2:2, 159-165.

- 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
     - mutual information
     - log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

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

Topic signature-based content selection with queries

Conroy, Schlesinger, and O'Leary 2006

- choose words that are informative either
  - by log-likelihood ratio (LLR)
  - or by appearing in the query

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

(could learn more complex weights)

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

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

Supervised content selection

- 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 sentence?)
  - length of sentence
  - word informativeness, cue phrases
  - cohesion
- Train
  - a binary classifier (put sentence in summary? yes or no)

- Problems:
  - hard to get labeled training data
  - alignment difficult
  - performance not better than unsupervised algorithms
- So in practice:
  - **Unsupervised content selection is more common**

Evaluating Summaries: ROUGE

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Lin and Hovy 2003

- Intrinsic metric for automatically evaluating summaries
  - Based on BLEU (a metric used for machine translation – precision-driven)
  - Not as good as human evaluation ("Did this answer the user's question?")
  - But much more convenient
- 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?

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

A ROUGE example:
Q: “What is water spinach?”

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

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

Query-focused Summarization

- Or complex question answering

Definition questions

Q: What is *water spinach?*

A: Water spinach (ipomoea aquatica) is a semi-aquatic leafy green plant with long hollow stems and spear- or heart-shaped leaves, widely grown throughout Asia as a leaf vegetable. The leaves and stems are often eaten stir-fried flavored with salt or in soups. Other common names include *morning glory vegetable, kangkong* (Malay), *rau muong (Viet.), ong choi (Cant.), and kong xin cai* (Mand.). It is not related to spinach, but is closely related to sweet potato and convolvulus.

Medical questions

Demner-Fushman and Lin (2007)

Q: In children with an acute febrile illness, what is the efficacy of single medication therapy with acetaminophen or ibuprofen in reducing fever?

A: Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses. (*PubMedID: 1621668, Evidence Strength: A*)

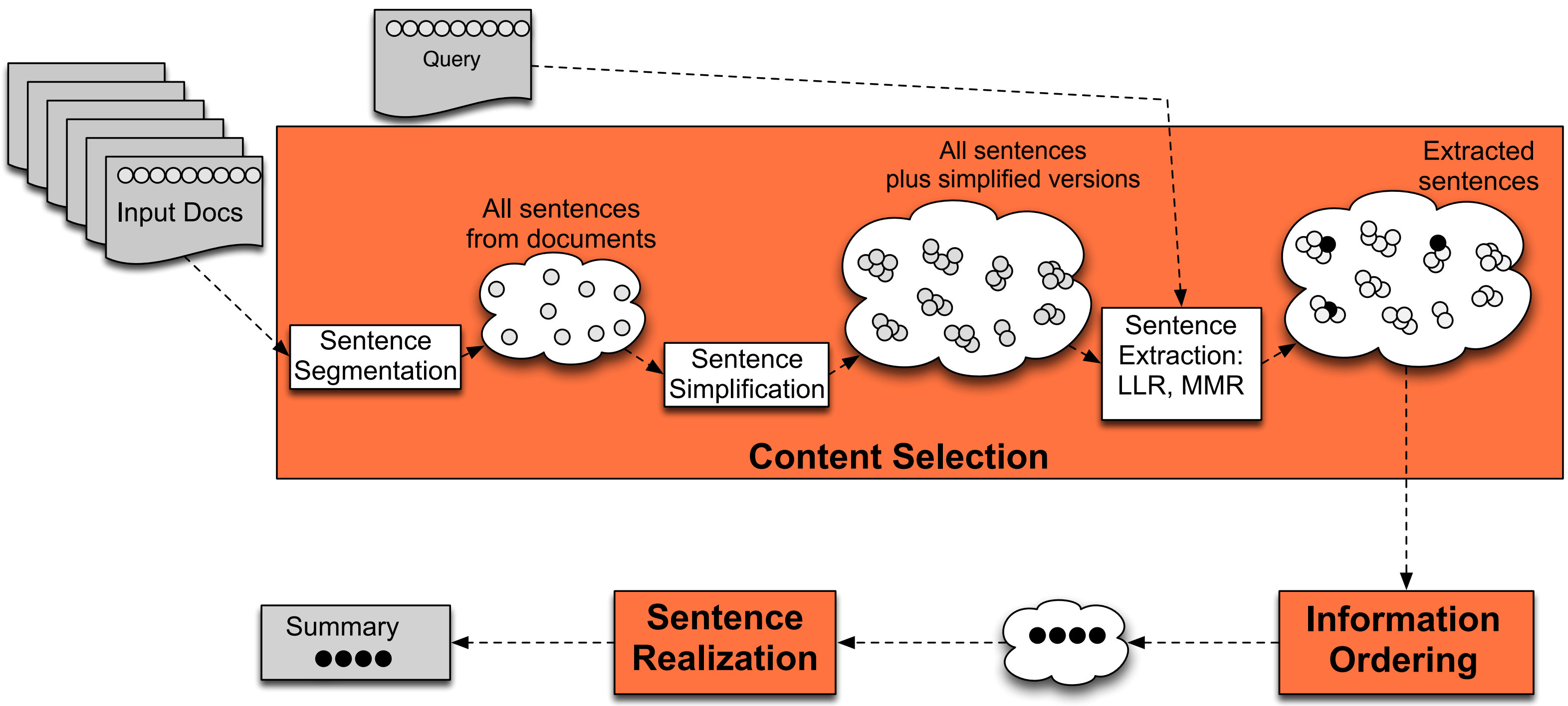
Other complex questions

1. How is compost made and used for gardening (including different types of compost, their uses, origins and benefits)?
2. What causes train wrecks and what can be done to prevent them?
3. Where have poachers endangered wildlife, what wildlife has been endangered and what steps have been taken to prevent poaching?
4. What has been the human toll in death or injury of tropical storms in recent years?

Answering harder questions: Query-focused multi-document summarization

- The (bottom-up) snippet method
  - Find a set of relevant documents
  - Extract informative sentences from the documents
  - Order and modify the sentences into an answer
- The (top-down) information extraction method
  - build specific answerers for different question types:
    - definition questions
    - biography questions
    - certain medical questions

Query-Focused Multi-Document Summarization



Simplifying sentences

Zajic et al. (2007), Conroy et al. (2006), Vanderwende et al. (2007)

Simplest method: parse sentences, use rules to decide which modifiers to prune
(more recently a wide variety of machine-learning methods)

appositives	Rajam, ~~28, an artist who was living at the time in Philadelphia~~, found the inspiration in the back of city magazines.
attribution clauses	Rebels agreed to talks with government officials, ~~international observers said Tuesday.~~
PPs without named entities	The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [~~PP to a sustainable number~~]
initial adverbials	"~~For example~~", "~~On the other hand~~", "~~As a matter of fact~~", "~~At this point~~"

Query-focused Multi-document Summarization Maximal Marginal Relevance (MMR)

Jaime Carbonell and Jade Goldstein, The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries, SIGIR-98

- An iterative method for content selection from multiple documents
- Iteratively (greedily) choose the best sentence to insert in the summary/answer so far:
  - Relevant: Maximally relevant to the user's query
    - high cosine similarity to the query
  - Novel: Minimally redundant with the summary/answer so far
    - low cosine similarity to the summary

$$\hat{s}_{MMR} = \max_{s \in D} \lambda sim(s,Q) - (1\text{-}\lambda) \max_{s \in S} sim(s,S)$$

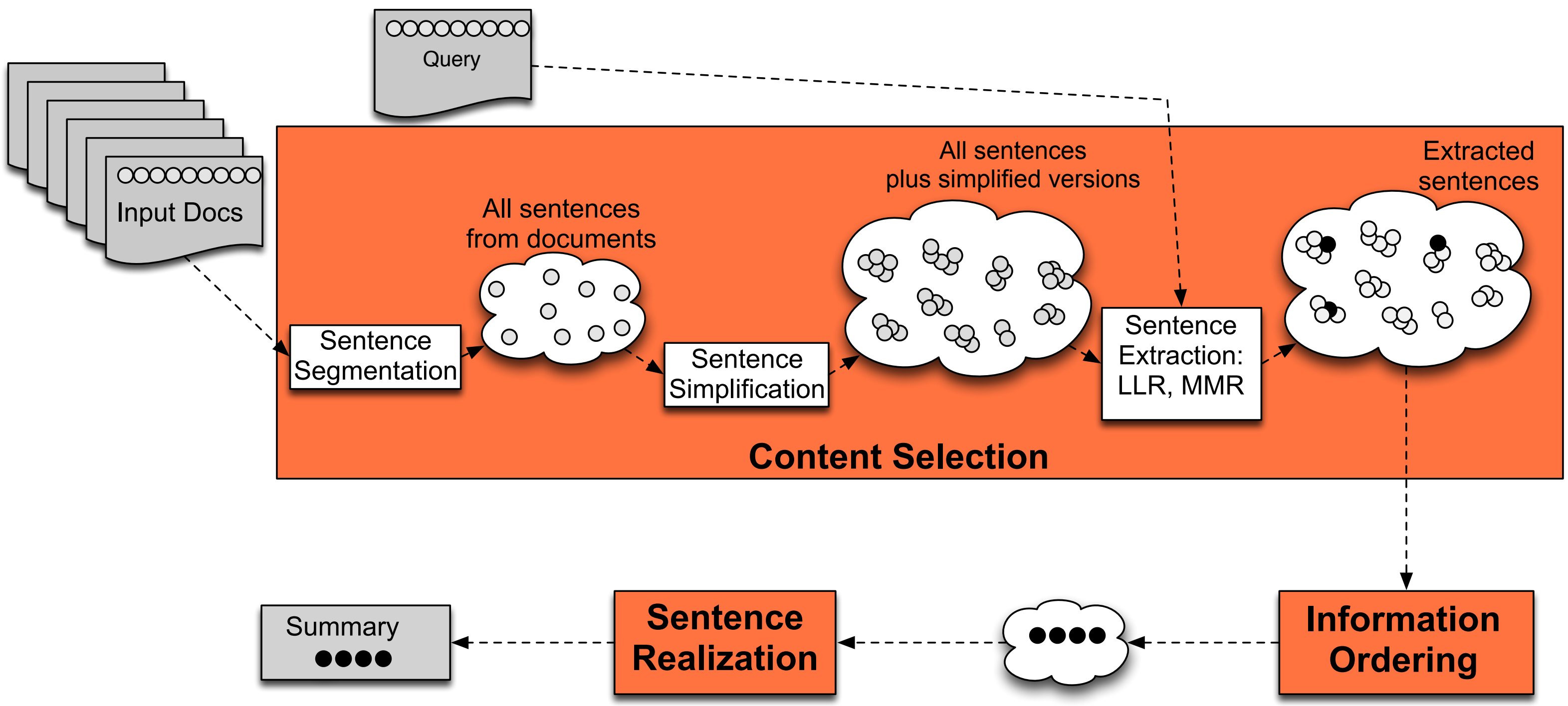
- Stop when desired length

LLR+MMR: Choosing informative yet non-redundant sentences

- One of many ways to combine the intuitions of LLR and MMR:

1. Score each sentence based on LLR (including query words)
2. Include the sentence with highest score in the summary.
3. Iteratively add into the summary high-scoring sentences that are not redundant with summary so far.

Query-Focused Multi-Document Summarization



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

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

Before

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.

After

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.